


计算机科学丛书

原书第2版

数据挖掘技术

市场营销、销售与客户关系管理领域应用

(美) Michael J. A. Berry　Gordon S. Linoff 著　别荣芳　尹静　邓六爱 译



WILEY

TIMELY. PRACTICAL. RELIABLE.

Data Mining Techniques

Second Edition

For Marketing, Sales, and Customer Relationship Management

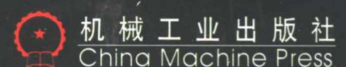


Michael J. A. Berry
Gordon S. Linoff

Data Mining Techniques

For Marketing, Sales, and Customer Relationship Management

Second Edition



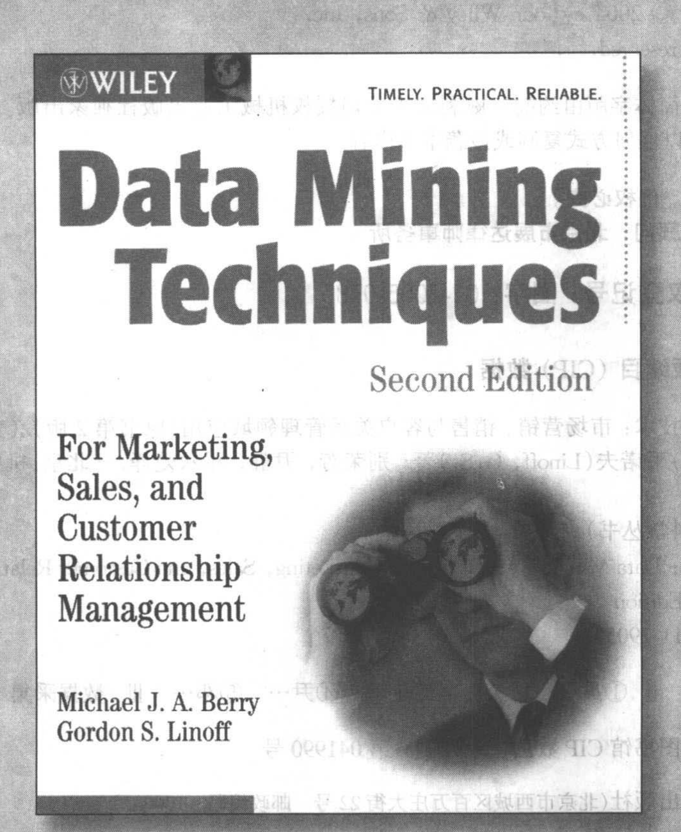
计算机科学丛书

原书第2版

数据挖掘技术

市场营销、销售与客户关系管理领域应用

（美）Michael J. A. Berry Gordon S. Linoff 著 别荣芳 尹静 邓六爱 译



Data Mining Techniques

For Marketing, Sales, and Customer Relationship Management

Second Edition



机械工业出版社
China Machine Press



本书是一本优秀的数据挖掘教材，全面而系统地介绍了数据挖掘的商业环境、数据挖掘技术及其在商业环境中的应用。

全书共18章，内容涵盖核心的数据挖掘技术，包括：决策树、神经网络、协同过滤、关联规则、链接分析、聚类和生存分析等。此外，还提供了数据挖掘最佳实践的概观、数据挖掘的最新进展和一些极具挑战性的研究课题，极具技术深度与广度。通过学习本书，读者不仅可以精通数据挖掘的整体结构和核心技术，还可以领略数据挖掘在销售和客户关系管理等方面的成功应用，为实践数据挖掘打下坚实的基础。

本书适合作为高等院校相关专业高年级本科生或研究生的教材或参考书，也适合当前和未来的数据挖掘实践者学习和参考。

Michael J. A. Berry, Gordon S. Linoff: Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Second Edition (ISBN: 0-471-47064-3)

Authorized translation from the English language edition published by John Wiley & Sons, Inc.

本书版权登记号：图字：01-2005-0785

图书在版编目（CIP）数据

数据挖掘技术：市场营销、销售与客户关系管理领域应用（原书第2版）/（美）贝瑞（Berry, M.J.A.），（美）莱诺夫（Linoff, G.S.）著；别荣芳，尹静，邓六爱译. －北京：机械工业出版社，2006.7

（计算机科学丛书）

书名原文：Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Second Edition

ISBN 7-111-19056-4

Ⅰ.数…　Ⅱ.①贝…　②莱…　③别…　④尹…　⑤邓…　Ⅲ.数据采集　Ⅳ.TP274

中国版本图书馆CIP数据核字（2006）第041990号

机械工业出版社（北京市西城区百万庄大街22号　邮政编码　100037）

责任编辑：朱起飞

北京京北制版印刷厂印刷·新华书店北京发行所发行

2006年7月第1版第1次印刷

184mm×260mm·26.75印张

定价：49.00元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

本社购书热线：（010）68326294

出版者的话

文艺复兴以降，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的传统，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭橥了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短、从业人员较少的现状下，美国等发达国家在其计算机科学发展的几十年间积淀的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章图文信息有限公司较早意识到“出版要为教育服务”。自1998年开始，华章公司就将工作重点放在了遴选、移译国外优秀教材上。经过几年的不懈努力，我们与Prentice Hall，Addison-Wesley，McGraw-Hill，Morgan Kaufmann等世界著名出版公司建立了良好的合作关系，从它们现有的数百种教材中甄选出Tanenbaum，Stroustrup，Kernighan，Jim Gray等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及庋藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力襄助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专程为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍，为进一步推广与发展打下了坚实的基础。

随着学科建设的初步完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都步入一个新的阶段。为此，华章公司将加大引进教材的力度，在“华章教育”的总规划之下出版三个系列的计算机教材：除“计算机科学丛书”之外，对影印版的教材，则单独开辟出“经典原版书库”；同时，引进全美通行的教学辅导书“Schaum's Outlines”系列组成“全美经典学习指导系列”。为了保证这三套丛书的权威性，同时也为了更好地为学校和老师们服务，华章公司聘请了中国科学院、北京大学、清华大学、国防科技大学、复旦大学、上海交通大学、南京大学、浙江大学、中国科技大学、哈尔滨工业大学、西安交通大学、中国人民大学、北京航空航天大学、北京邮电大学、中山大学、解放军理工大学、郑州大学、湖北工学院、中国国家信息安全测评认证中心等国内重点大学和科研机构在计算机的各个领域的著名学者组成“专家指导委员会”，为我们提供选题意见和出版监督。

这三套丛书是响应教育部提出的使用外版教材的号召，为国内高校的计算机及相关专业

的教学度身订造的。其中许多教材均已为M. I. T.，Stanford，U.C. Berkeley，C. M. U. 等世界名牌大学所采用。不仅涵盖了程序设计、数据结构、操作系统、计算机体系结构、数据库、编译原理、软件工程、图形学、通信与网络、离散数学等国内大学计算机专业普遍开设的核心课程，而且各具特色——有的出自语言设计者之手、有的历经三十年而不衰、有的已被全世界的几百所高校采用。在这些圆熟通博的名师大作的指引之下，读者必将在计算机科学的宫殿中由登堂而入室。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证，但我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。教材的出版只是我们的后续服务的起点。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下:

电子邮件: hzjsj@hzbook.com
联系电话: （010）68995264
联系地址: 北京市西城区百万庄南街1号
邮政编码: 100037

译 者 序

随着数据库技术的应用越来越普及，人们逐渐陷入了“数据丰富，知识贫乏”的尴尬境地，因为大量数据淹没了数据中隐含的模式和有益信息。于是，致力于摆脱这一困境的数据挖掘技术从20世纪90年代起步并得到迅速发展。数据挖掘技术是数据库研究、开发和应用中最活跃的分支之一，是一种基于机器学习、统计分析等多种学科的计算机技术，能够有效地帮助人们将海量数据资源转换为有用的知识和信息，进而帮助人们科学地做出决策。

本书是数据挖掘领域的巨著，多年以来，在数据挖掘领域的地位始终无可替代，其内容也随数据挖掘技术的发展演化而不断更新。本书最早的版本是1997年出版的，补充修订后于2004年出版第2版。新版中减少了与商业相关的素材，增加了更多的技术素材，并加入了作者近年来的最新研究成果和见解，比如：关于数据挖掘在营销和客户关系管理方面的应用、基本统计学技术的使用、生存分析和为挖掘准备数据等内容。基于存储的推理增加了以最近邻技术为基础的协同过滤方法，从而在技术和应用两方面更加全面、系统地介绍了数据挖掘的商业环境、数据挖掘技术及其在商业环境中的应用。

本书共有18章，内容涵盖了核心的数据挖掘技术，包括：决策树、神经网络、协同过滤、关联规则、链接分析、聚类和生存分析等。此外，还提供了数据挖掘最佳实践的概观、数据挖掘的最新进展和一些极具挑战性的研究课题，其技术深度与广度举世公认。作者注重实效，对每类问题均提供代表性算法，以亲身经历的商业案例为实例，给出每一技术具体的应用法则。通过学习本书，读者不仅可以精通数据挖掘的整体结构和核心技术，还可以领略数据挖掘在营销、销售和客户关系管理等方面的成功应用，为实践数据挖掘打下坚实的理论和应用基础。

本书的目标读者是当前和未来的数据挖掘实践者，可以作为相关专业高年级本科生的选修课教材，特别适合作为研究生的专业课教材。本书用生活实例开头，引出基本概念，同时提供大量真正的商业环境实例。因此，对于从事数据挖掘应用的读者来说，是一本必备的参考书。本书的网站还有一些推荐读物和练习，所以对于初学者来说，也是一本可读性极佳、适于循序渐进地学习数据挖掘的首选教科书。

本书主要由别荣芳、尹静和邓六爱三位翻译完成。全书由别荣芳统一审校。孙运传参与了部分审校工作。在翻译过程中，译者发现一些错误和疑似错误之处，在译文中对一般拼写错误和明显笔误均未作说明而直接进行了校正，其他错误则在相应页的脚注中给出了说明。

由于时间仓促，加上本书涉及诸多实际应用领域，原作中方言俚语和非信息技术专业词汇较多，翻译内容难免存在疏漏和不足，敬请读者谅解并批评指正。

译 者
2006.6

致 谢

非常幸运的是，我们周围有很多天才的数据挖掘专家，因此首先要感谢在 Data Miners 公司的同事，从他们那里我们学到了很多。他们是：Will Potts、Dorian Pyle 和 Brij Masand。还有许多曾经与我们密切合作的客户，我们也把他们视为同事：Harrison Sohmer 和 Stuart E. Ward，Ⅲ。编辑 Bob Elliott、编辑助理 Erica Weinstein 和责任编辑 Emilie Herman 帮助我们把握进度，并保持风格一致。毕业于麻省理工学院的 Lauren McCann，在 Data Miners 公司实习期间，准备了在很多例子中使用的人口普查数据，并创建了一些图表。

我们还要感谢过去多年来在数据挖掘方面与我们共事的所有人。我们从每个人那里学到了很多。那些数据挖掘方案对本书第 2 版有影响的人包括：

Al Fan
Alan Parker
Anne Milley
Brian Guscott
Bruce Rylander
Corina Cortes
Daryl Berry
Daryl Pregibon
Doug Newell
Ed Freeman
Erin McCarthy
Herb Edelstein
Jill Holtz
Joan Forrester
John Wallace
Josh Goff
Karen Kennedy
Kurt Thearling
Lynne Brennen
Mark Smith
Mateus Kehder
Michael Patrick
Nick Gagliardo
Nick Radcliffe
Patrick Surry
Ronny Kohavi
Sheridan Young
Susan Hunt Stevens
Ted Browne
Terri Kowalchuk
Victor Lo
Yasmin Namini
Zai Ying Huang

当然，我们仍然要感谢在第 1 版曾经感谢的人们：

Bob Flynn
Bryan McNeely
Claire Budden
David Isaac
David Waltz
Dena d' Ebin
Diana Lin
Don Peppers
Ed Horton
Edward Ewen
Fred Chapman
Gary Drescher
Gregory Lampshire
Janet Smith
Jerry Modes
Jim Flynn
Kamran Parsaye
Karen Stewart
Larry Bookman
Larry Scroggins
Lars Rohrberg
Lounette Dyer
Marc Goodman
Marc Reifeis
Marge Sherold
Mario Bourgoin
Prof. Michael Jordan
Patsy Campbell
Paul Becker
Paul Berry
Rakesh Agrawal
Ric Amari
Rich Cohen
Robert Groth
Robert Utzschnieder
Roland Pesch
Stephen Smith
Sue Osterfelt
Susan Buchanan
Syamala Srinivasan
Wei-Xing Ho
William Petefish
Yvonne McCollin

前　言

本书第1版于1997年面世。该书实际上开始于1996年，当时我和Gordon在为国家银行（NationsBank）（现在是美国银行，Bank of America）设计一天的数据挖掘研讨班。NationsBank的一位副总裁Sue Osterfelt（她还与Bill Inmon合著有一本关于数据库应用的图书）使我们深信，研讨班的材料应该整理成一本书。她把Jon Wiley & Sons公司的编辑Bob Elliott介绍给我们，在我们还没来得及仔细考虑这件事情之前，就签了一份合同。

我们两个人以前从未写过书，前面几章的草稿清楚地说明了这一点。感谢Bob的帮助，我们取得了很大的进步，最终版本仍然是相当令人骄傲的。毫不夸张地说，这一经历改变了我们的生活：第一是占用了应该散步的每一小时，甚至是应该睡觉的时间；其次，更肯定地说，提供了我们创建的Data Miners咨询公司的基础。本书第1版已经成为数据挖掘的一本标准教材，后续著作包括：*Mastering Data Mining*和*Mining the Web*。

那么为什么要进行修订呢？自从第1版出版以来，数据挖掘界发生了很大的变化。例如：那时候，Amazon.com才刚刚出现；美国移动电话呼叫费用平均为每分钟50美分，不超过25%的美国人拥有移动电话；KDD数据挖掘会议才举办了第二届。我们的理解也改变了很多。尽管其中的大部分核心算法仍然保持不变，但是算法嵌入的软件、应用算法的数据库以及用于解决的商业问题都有所增长和演化。

即使技术界和商业界保持不变，我们也希望更新本书第1版，因为在其间的几年，我们又学到了很多。做咨询的一大乐趣就是时刻面对新思想、新问题和新的解决方案。我们并不比当年写第1版的时候更聪明，但确实经验更丰富，而且我们的写作经验也更丰富了。稍微浏览一下本书内容目录就可以发现，我们减少了很多与商业相关的材料，而增加了更多的技术材料。另外，把一些商业材料融汇到技术章节中，因此使数据挖掘技术得以在商业环境中来讨论，希望这样可以使读者更容易领会到如何把技术应用到自己的商业问题。

我们还注意到，许多商业学校的课程使用本书作为教材。尽管我们并没有把本书写成一本教科书，在第2版中，我们努力使它可以用作教材，书中提供了大量基于公开可用的数据的实例，诸如美国的人口普查数据，在配套网站www.data-miners.com/companion中有推荐阅读材料和建议的练习。

全书仍然分为三个部分，第一部分讲述数据挖掘的商业环境。开篇章节给出了数据挖掘的简介，解释数据挖掘可以用来干什么，并且为什么需要数据挖掘。第2章介绍数据挖掘的良性循环，这是一个持续不断的过程，通过这个过程，数据挖掘将数据转变为指导行动的信息，反过来创造了更多的信息和更多的学习机会。第3章是数据挖掘的方法论和最佳实践的拓展讨论，该章比书中任何其他一章更得益于我们写第一本书以来的经历，这里介绍的方法论基于我们曾经参与的成功案例而设计。第4章在第1版中根本没有相应的部分，是关于数据挖掘在营销和客户关系管理中的应用，也正是我们现在所从事的领域。

第二部分讲解数据挖掘本身的技术内容，包含第1版描述的所有技术，但是重新进行了调整，对各种描述进行了重写，比第1版更清晰、更准确。但仍然保留了第1版的风格，即可能的地方都使用非技术语言。

除了包含第1版涵盖的7种技术：决策树、神经网络、协同过滤、关联规则、链接分析、聚类和生存分析之外，还增加了使用基本的统计学技术以及生存分析的新章节。生存分析是一项广泛应用的技术，从医学界的少量样本和连续的时间测量，到营销数据中发现的大量样本和离散时间度量，都可应用。基于存储的推理一章还包括以最近邻技术为基础的协同过滤方法，作为产生推荐的方式，已经为广大Web零售商所熟知。

第三部分讲述在商业环境中使用技术的方法，其中有一章关于在数据中发现客户，另一章关于数据挖掘和数据仓库的关系，还有一章关于数据挖掘环境（公司环境和技术环境两个方面），最后一章关于在公司中应用数据挖掘。该部分新增加了一章，介绍为数据挖掘准备数据。这是一个极其重要的话题，因为很多数据挖掘者反映，在典型的数据挖掘工程中，转换数据通常需要花费大多数的时间。

和第1版一样，本书仍然针对当前和未来的数据挖掘实践者。既不是为软件开发者提供如何实现各种数据挖掘算法的细节指导，也不是为了使研究人员改进那些算法。有关思想以非技术的语言给出，尽可能少地使用数学公式和艰涩的术语。每一种数据挖掘技术都在真实的商业环境中展示，给出大量来自商业环境的实例。简而言之，我们努力把本书写成打算开始数据挖掘生涯的技术人员喜欢读的一本书。

Michael J. A. Berry

2003年10月

目　录

第1章　数据挖掘的缘起和内容

在本书第1版中，第1章的第一句就写到："马萨诸塞州萨默维尔市，本书作者之一的故乡……"，接着讲述了那个镇上的两个小店和他们如何与客户形成学习关系（learning relationship）的故事。该章描述了梳小辫的小女孩和给她梳辫子的人的关系，在其间的几年中，这个小女孩已经长大成人，离开小镇，也不再梳着小辫，她的父亲也搬到附近的剑桥居住。但是有一件事情没变，作者仍然是Wine Cask商店的忠实客户。正是在这个小店，同样忠诚的一些客户在1978年将便宜的阿尔及利亚红酒介绍给他，后来介绍给他法国的葡萄酒产区，现在正帮他开发意大利和德国的酒源。

25年后，他们仍然有一位忠实的客户，这并非偶然。在Wine Cask商店的Dan和Steve了解他们的客户的口味和可承受价位，当有客户询问时，他们的回答除了基于本店库存外，还有因日积月累而得到的有关该顾客口味和经济能力方面的信息。

Wine Cask商店的人掌握很多有关葡萄酒的知识，尽管这种知识是很多人来这里买酒而不是去大的折扣酒店的原因之一，但是他们对每个客户的详细了解才是客户持续购买的主要原因。也许可以在大街对面开另一个酒店，同样雇用一批品酒专家，但是要达到对客户了解程度具有同样水平至少需要几个月甚至数年时间。

经营好的小商店自然与他们的客户形成学习关系。久而久之，他们对客户的了解越来越多，然后用这种了解更好地为客户服务，结果不仅获得忠实的客户，还盈利颇丰。拥有数十万乃至上百万客户的大公司，难以形成与每个客户的密切关系，这些公司必须依赖其他方法形成与客户的学习关系。特别是，他们必须充分利用自己拥有的大量东西，那就是几乎每笔客户交易所产生的数据。本书将要讲述的就是如何把客户数据转换为客户知识的分析技术。

1.1　分析客户关系管理系统

人们普遍认为，任何规模的公司都需要学会效仿那些以服务为本的小企业的成功之处——与客户建立一对一的关系。客户关系管理（customer relationship management，CRM）系统是很多书和会议中广泛讨论的主题，从引导追踪软件到调用中心软件的外围管理软件都被称为客户关系管理工具。本书主要关注的是数据挖掘（data mining）在提高公司与客户形成学习关系的能力，进而改善客户关系管理中所起的作用。

在任何行业，有远见的公司正在向着下面的目标努力：努力了解每个客户个体，并且利用这种了解使客户选择与他们进行商业活动，而不是选择他们的竞争对手。这些公司也正在学习认识每个客户的价值，进而知道哪些人值得投入资金和精力来保持联系，哪些人可以放弃。从重视广泛的市场到重视客户个体的这种转变，需要整个企业在市场、销售和客户支持等方面适应这种转变。

对大多数公司来说，围绕客户关系建立商业活动是一种全新的变革。银行一贯关注如何保持存款应付利息和贷款应收利息的差额，电信公司关注网络内通话连接，保险公司关注处理理赔和投资管理。仅使用数据挖掘并不足以把一个注重产品的组织转变为以客户为中心的组织。如果管理者的奖金基于新物品的季度销售数量而不是小部件的销售数量，一个建议给

某个客户提供一个小部件而不是一件新物品的数据挖掘结果极容易被忽略，尽管也许后者盈利更多。

狭义地讲，数据挖掘是一系列工具和技术的集合，是支持以客户为中心的组织需要的多项技术之一；广义地讲，数据挖掘是一种态度，它表明商业活动应该基于认知，分析获得的决策比没有任何分析所得的决策好得多，经过测算的结果更利于商业盈利。数据挖掘还是应用这些工具和技术的过程和方法论。为进行有效的挖掘，分析客户关系管理系统的其他要求也必须到位。为了与其客户形成学习关系，公司必须做到：

- 注意客户正在做什么
- 记住公司及其客户曾经做过什么
- 从记住的信息学习
- 按照获得的知识进行商业活动使顾客更加受益

本书的目标是上述第三个方面，也即从过去发生的事情中学习，这种学习不可能凭空进行。必须依靠交易处理（transaction processing）系统收集客户数据，用数据仓库存储客户历史行为信息，使用数据挖掘把历史数据转变成未来行动计划，然后通过某种客户关系策略将这一计划付诸实施。

1.1.1 交易处理系统的作用

小企业通过注意客户的需求，记住客户的喜好，从过去的交流中学习如何在未来更好地服务他们，由此建立与客户的关系。但是对于大多数雇员从来不与客户交流的大公司来说，如何完成类似的事情呢？在这种公司中即使有一些客户交流，也可能仅仅是与销售职员或不知名字的客服中心的员工进行交流，那么，公司怎么可能注意到或记住这些信息，并且从这种交流中获取信息呢？什么东西能够替代可以识别客户姓名、面孔、声音，记住客户的习惯和喜好的独特的创造性直觉呢？

一句话，没有东西可以代替它！但这不代表我们不可以做尝试。通过灵活运用信息技术，即使是最大的公司也可获得惊人的相近结果！在大商业公司，注意客户的行为这一步已经高度自动化，交易处理系统无处不在，收集几乎所有的数据：自动售货机、电话交换机、网络服务器和售点扫描仪等生成的数据，都是数据挖掘的主要素材。

目前，每天的生活都可产生一系列的交易记录。当拿起电话从 L.L.Bean 预订一只皮划艇桨或者从 Victoria's Secret 定制一个缎纹文胸，市话公司就生成详细电话记录，显示呼叫时间、呼叫电话号码以及被叫长途电话公司等。在长途电话公司，也会生成类似的记录，包括持续通话时间和使用的交换机中的具体路由线路。这些数据连同个人账号信息、姓名和地址等其他记录产生一个账单。订购公司也会记录你的呼叫，连同预订项信息以及对一些推销商品的反应。当接听电话的销售服务代表询问你的信用卡号码以及交付期限时，信息很快转入转账的信用卡验证系统，这样又生成了一条记录。然后转账业务抵达发行信用卡的银行，出现在下个月的银行账单中。当订单连同商品号码、型号和颜色进入订单系统，在付账系统和库存控制系统中将产生另外的记录。几小时后，你的订单又会在 UPS 或者 FedEx 的计算机系统中产生交易记录，它们在你家和公司仓库之间进行多次扫描，可以使你通过检查邮递公司的网页来方便地追踪所订购的物品。

这些交易记录不是专门为数据挖掘生成的，而是公司的运作需要。然而所有记录均包含

重要的客户信息，并且可以被成功挖掘。电话公司利用详细通话记录发现哪些居民的电话号码类似商用，从而对在家中进行商业活动的人推出特殊服务。订购公司利用历史订单判断哪些客户应该包含在哪种未来邮件名录中，以 Victoria' s Secret 为例，它可以发现哪种模式可达到最好销售状况。联邦快递公司（federal express）在 UPS 员工的一次罢工期间通过对客户运送模式的改变来估算它在客户货物运送业务的份额。超级市场运用销售点数据来决定对哪些用户使用何种优惠券。网络零售商使用过去的购买情况来确定当客户浏览网站时应该展示哪些商品。

这些交易系统是客户接触点，从那里客户行为信息首次进入公司，因而，它们也是公司的眼睛和耳朵（也许是鼻子、舌头或者手指）。

1.1.2 数据仓库的作用

关注客户的公司把与每位客户或者潜在客户的每条互动记录视为一次学习机会，这些互动包括打给客户服务中心的每次电话、每一笔销售点交易、每一个订单、对公司网站的每一次访问。但是学习不只是收集数据，事实上，很多公司收集数千亿字节甚至百万亿字节的客户数据，却没有获取任何有用的信息。获取数据是某些运作的需要，如库存控制或者付账，一旦达到目的，数据将被搁置在磁盘或者磁带上，或者被丢弃。

为了研究客户状况，首先必须将从各种渠道收集的数据，如付账记录、扫描数据、登记表格、申请、电话记录、优惠券兑换情况和调查表等，用某种一致和有效的方式组织起来，这就是数据仓库（data warehousing）。数据仓库能使公司记住自己的客户的情况。

提示： 客户模式随着时间的推移而日趋清晰。数据仓库需要提供精确的历史数据，以便通过数据挖掘得到更可信的趋势。

数据仓库的最重要的特征之一是随时间变化追踪客户行为的能力。客户关系管理系统感兴趣的很多模式只能随时间日趋显现：使用趋势是上升还是下降？客户回头的频繁程度如何？客户更倾向哪种方式？客户对哪种促销形式有回应？

多年以前，当一个大的目录零售商（通过商品目录册订购进行销售，catalog retailer）首次保存了一年以上的历史投递目录以及客户的响应后，发现了保持客户历史行为数据的重要性。他们发现，一些客户只有在圣诞节期间才从目录中预订。利用这部分用户的信息，他们决定做些尝试，提出一种方法在一年的其他时间里刺激客户下订单的兴趣，或者通过在那段时间不给这些客户投递来增加客户总体响应率。虽然没有作进一步的试验，还不知道哪个方式正确，但假如没有历史数据的帮助，就永远不知道考虑这个问题。

原始数据产生于操作系统并存储其中，但好的数据仓库提供了一种更为友好的访问从交易数据中提取的信息的方法。理想的情况是，数据仓库中来源于多个数据源的数据通过清理、合并、与某个客户关联，汇总成各种有益的形式。现实情况中通常达不到这种理想状态，但是公司的数据仓库仍然是分析客户关系管理的最重要的数据源。

1.1.3 数据挖掘的作用

数据仓库为企业提供数据存储，但是非智能的存储毫无用处。智能允许我们梳理存储信息，注意某些模式，设计规则，提出新思想，解决关键问题，预测未来趋势。本书阐述了为数据仓库增加智能特性的工具和技术，这些技术使得使用客户数据更进一步了解客户成为

可能。

谁可能仍然是忠实的客户？谁可能逃掉？什么产品应该以何种定位面世？是什么决定某个客户是否对某种产品做出回应？哪种电话销售方式最适合某个客户？下一个分支机构应该设在哪里？某个客户需要的下一种产品或者服务是什么？类似这些问题的答案就隐藏在公司数据中，需要强有力的数据挖掘工具才能找到这些答案。

用于客户关系管理的数据挖掘的核心思想是过去的数据包含对未来有用的信息。因为在公司数据中获取的客户行为不是杂乱无章的，而是反映了客户的不同需要、倾向、嗜好以及处理方式。数据挖掘的目标是从历史数据找寻不同的模式，这些模式清楚反映了这些需要、倾向和嗜好。事实上，由于这些模式并不总是很清晰，客户给出的信号有噪声，从而使这项任务变得很困难。从噪声中分离信号，从看似随意的变化中识别主要的模式是数据挖掘的一项重要任务。

本书涵盖了几乎所有重要的数据挖掘技术，以及每项技术在客户关系管理环境下的优缺点。

1.1.4 客户关系管理策略的作用

为达到有效性，数据挖掘必须在事情发生的环境内进行，这种环境允许企业从获得的知识来改变其行为。如果没有人能够给手机用户提供更适合的价位套餐，知道他可能因为选择了错误的手机套餐而打算退出也是没有用的。数据挖掘应该嵌入整体客户关系策略，通过这个策略可以清楚地知道根据从数据挖掘中获取的知识所应采取的动作。一旦确定低价值的用户，应该如何对待他们？有哪些计划可以刺激他们的消费从而增加他们的价值？或者降低为他们服务的成本更有意义？如果某些渠道可以带来更有利可图的客户，相应的资源该如何转向这些渠道？

数据挖掘是一个工具。像其他工具一样，只知道它如何工作是不够的，还必须了解如何应用它。

数据挖掘建议和业务决策

本段稍微详细地探索正文中的例子。手机服务提供商的消耗分析常常显示：当用户的呼叫模式与其价位套餐不匹配时，用户可能取消业务。用户使用电话的时间超出计划时，超出部分通常要付很高的价格；而没有用完全部时间的用户，剩余的时间部分仍会按分钟数收费，这样的用户就可能被其他竞争者提供的更便宜的套餐吸引走。

这种结果表明，应该预先积极地做好工作，使客户使用合适的价位套餐。但这不是一个简单的决定能够完成的。只要这些使用不合适价位套餐的客户不退出，任其自然，公司从他们那里可以赚取更高的利润。进一步分析，也许其中一部分客户对价格不敏感，他们也许会安于现状，但是任何小的动作都有可能给客户提供退出的机会。或许一个大小适中的测试可以解决这些问题，数据挖掘有助于做出更可行的决定。它可以为需要做的测试提出些建议，但最终由企业做出决策。

1.2 什么是数据挖掘

顾名思义，数据挖掘是探查和分析大量数据以发现有意义的模式和规则的过程。对于本

书，我们假设数据挖掘的目标是允许公司通过对客户的更好了解来改善其市场、销售和客户支持运作。然而应该说明的是，本书描述的技术和工具同样适用于其他领域，从法律的实施到射电天文学，以及医药和工业过程控制等。

事实上，几乎没有哪个数据挖掘算法是专为商业应用而发明的。商用数据挖掘工具从统计学、计算机科学和机器学习研究等方面借鉴了很多技术。究竟选择哪些数据挖掘技术的组合以应用于某个具体情况，取决于数据挖掘任务自身、可用数据的种类以及数据挖掘人员的偏好和技巧。

数据挖掘分为定向和非定向两类。定向数据挖掘的目的是解释或者分类某个特殊的目标域，如收入或者反馈。非定向数据挖掘的目的是在不预设目标域或确定类的前提下，找出在批量数据间的模式或者相似性。这两种类型都将在后面的章节介绍。

数据挖掘与模型构造密切相关。模型就是把一组输入关联到一个特定输出的一个算法或者规则集，这里的输入通常是公司数据库字段的形式。回归（regression）、神经网络（neural network）、决策树（decision tree）和本书中讨论的其他大部分技术都是构造模型的技术。在适当的情况下，通过解释某种特定结果（如下订单或者未付账等）如何与已知事实相关，模型可以给出更好的理解。模型也可用来产生得分（score）。得分是以一个简单的数值来表述模型输出的一种方式。得分可用于将客户排序，从最忠诚到最不忠诚，从最可能响应到最少响应，从最可能拖欠贷款到最不可能拖欠贷款等。

数据挖掘过程有时也称为*知识发现*，或者数据库中的*知识发现*（knowledge discovery in databases，KDD）。我们更倾向于认为它是*知识创造*。

1.3 数据挖掘可以完成哪些工作

很多智能的、经济的以及商业利益问题可用短语表示为如下6类任务：

- 分类（classification）
- 估计（estimation）
- 预测（prediction）
- 关联分组（affinity grouping）或关联规则（association rule）
- 聚类（clustering）
- 描述和建立简档（description and profiling）

其中，前三项是定向数据挖掘的例子，目的是发现特定目标变量的值。关联分组和聚类是非定向挖掘的任务，目的是在不限定特定目标变量的情况下揭示数据的结构。建立简档可能是定向的，也可能是非定向数据挖掘任务。

1.3.1 分类

分类是最常见的数据挖掘任务之一，它似乎是人类的规则。为了理解并与周围环境交流，我们每天都在归类、分类以及分级。我们把生物分为门、种和纲，物质分解到不同元素，犬分为品种，人分种族，牛排和枫蜜分为USDA等级。

分类包括考察一类新出现的对象的特征，并归类到已定义类中。分类的对象通常表示为数据库表或者文件中的记录，分类工作包括向数据库添加一个新列，并给出某种分类代码。

分类工作首先要有一个清晰定义的类，还要有一系列已分类实例。分类过程实际上是先

建立某种模型，然后将其用于对未分类数据进行分类。

本书中已讲过的分类工作的例子包括：

- 将信用卡申请者分为低、中、高风险
- 选择在网页上展示的内容
- 确定哪些电话号码与传真机相连
- 发现欺骗性理赔申请
- 基于工种描述文本，指定行业代码和工种设计

所有这些例子中的类都是有限的，我们期望能够把新对象归入其中的某一个类中。决策树（在第6章讨论）和最近邻技术（在第8章讨论）都能很好地用于分类。神经网络（在第7章讨论）和链接分析（在第10章讨论）也是在某些情况下对分类有用的方法。

1.3.2 估计

分类给出的结果是离散的：是或否，是麻疹、风疹还是水豆。而估计则是处理连续值结果。输入一组数据，估计给出一个未知连续变量的值，如收入、高度或者信用卡的余额。

实际上，估计经常用于分类任务。如果一个信用卡公司希望向滑雪靴制造商出售账单信封封面广告空间，它可能建立的分类模型是把持卡人分为滑雪者或者非滑雪者两种。另一种方法是建造模型，对每个持卡人给以“滑雪倾向值分”，得分可以是0到1的数值，表示持卡人成为滑雪者的可能性。这样分类任务变为建立阈值得分，任何一位得分超过阈值的人被划为滑雪者，而低于这个值的人被认为是非滑雪者。

估计方法的优势是个人记录可以按照估计值排序。这一点的重要性可以从下面的例子中看出，假如滑雪靴制造公司打算投递50万封信件，如果确定有150万滑雪者，使用分类方法，它也许会简单地将广告寄给随意从这150万人中抽出的50万人。但是按照持卡人的“滑雪倾向值分”，公司可以把广告寄给最有可能的50万位候选者。

估计任务的例子还包括：

- 估计一个家庭的孩子数目
- 估计一个家庭的总收入
- 估计客户的寿命值
- 估计某人对余额转移诱惑的回应的可能性

回归模型（regression model，在第5章讨论）和神经网络（在第7章讨论）都非常适合估计任务。如果目的是估计一个事件的时间（如客户停止时间），生存分析（survival analysis，见第12章）也非常适合估计任务。

1.3.3 预测

预测与分类和估计一样，但其中记录的分类依据是一些预测的未来行为或者估计的未来值。在预测任务中，检验分类准确度的惟一方法是等待和观察。把预测从分类和估计中分离为单独的任务，主要是由于在预测建模时，存在其他关于输入变元的时序关系或者目标变元的预测问题。

所有用于分类和估计的技术均可稍加改变后用于预测，这种改变是利用训练样本中那些

已知的历史数据验证样本中要预测的变量值，这些变量值在训练样本中是已知值。历史数据用于构造模型，以解释当前观察到的行为。当这个模型应用于当前的输入，给出的结果就是对未来行为的预测。

本书中已讨论的数据挖掘技术涉及的预测任务例子包括：

- 预测当信用卡潜在用户收到转账单后，可能转账的额度
- 预测哪些客户在 6 个月之内可能离开
- 预测哪些电话用户会预订增值服务，例如三方通话或者声音邮件

只要训练数据以适当的形式存在，本书讨论的数据挖掘技术都可以应用于预测。选择哪种技术取决于输入数据的本质、预测数值的类型和预测解释的重要性。

1.3.4 关联分组或关联规则

关联分组的任务是确定哪些事情应该分在一起。原型例子是购物篮分析的核心任务，即在超市的购物车中哪些物品会放在一起。零售连锁店可以使用关联分组来计划商店货架或目录上的物品放置位置，以便把经常被一起购买的物品放在一起。

关联分组也可以用于确认交叉销售的机率，设计吸引人的产品或服务包（组）。

关联分组是由数据产生规则的一个简单方法。如果猫粮和小猫窝两种物品经常放在一起，我们可以产生两条关联规则：

- 买猫粮的人购买小猫窝的可能性为 P1；
- 买小猫窝的人购买猫粮的可能性为 P2。

关联规则将在第 9 章详细讨论。

1.3.5 聚类

聚类是把各不相同的个体分割为有更多相似性的子群或者簇的工作。聚类与分类的区别在于聚类不依赖于预先定义的类，而分类是以训练预分类样本构建的模型为基础，把每条记录分配到一个预定义的类中。

在聚类中，没有预定义的类和样本。记录完全依靠其自相似性被归为一类。如果簇有什么意义的话，结果也完全由使用者确定赋予该簇何种意义。不同症状集合也许代表不同的疾病，客户属性簇也许表示不同的市场份额。

聚类通常作为一些其他形式的数据挖掘或建模的前奏。例如，聚类通常作为市场分割的第一步，不是对“客户对哪些促销反应最好”提出一个统一的适合所有人的标准，而是首先将客户划分为簇，即划分为有相似购物习惯的人群，然后提问对每个簇哪种促销反应最好。聚类将在第 11 章中进行详细讨论，第 7 章讨论另一个有时用于聚类的技术——自组织图。

1.3.6 建立简档

数据挖掘的目的有时仅仅是描述在繁杂的数据库中正在进行的事件，在某种程度上加强我们对当前生成数据的人、产品或者进程的理解。一个好的行为描述经常也是对行为本身的一种解释，至少提示从哪里着手寻找解释。“支持民主党的女性在数量上大于男性”这样一个简单的描述就是美国政治中著名的性别差异的一个例子，它引起了大众广泛的兴趣，也导致了新闻记者、社会学家、经济学家和政治科学家的进一步研究，更不用说想进入政府机关

的候选人了。

决策树（第6章讨论）是对一个与特殊目标相关的客户（或任何其他事情）建立简档的强有力的工具。关联规则（第9章讨论）和聚类（第11章讨论）也可用于建立简档。

1.4 为什么现在研究

本书描述的多数数据挖掘技术早已经存在，至少作为学术算法已经存在数年或者数十年。然而，仅仅在过去的十几年中，商业数据挖掘才大规模地流行。这应归于下述几个因素的共同结果：

- 数据正在生成
- 数据正在形成数据仓库
- 计算能力足以承受
- 客户关系管理的兴趣增强
- 商业数据挖掘软件产品已经易于使用

下面依次考察上述的每一个因素。

1.4.1 数据正在生成

当存在大量数据时，数据挖掘最有意义。事实上，多数数据挖掘算法需要大量数据来建立或训练模型，以便进行分类、预测、估计或其他数据挖掘任务。

包括远程通信和信用卡公司在内的一些企业，已经与客户产生了一个自动化的交互关系，生成大量交易记录。但是仅仅到最近，日常生活自动化才变得普遍深入。现今，超级市场销售点扫描器、自动售货机、信用卡和借记卡、按次计费电视、在线购物、资金电子转账、自动化的订单处理和电子售票等类似手段的兴起意味着数据正在以空前的速度产生和收集。

1.4.2 数据正在形成数据仓库

数据不但已经大量产生，而且正在越来越频繁地由运作账单、（旅馆房间等）预订、索赔处理、订单系统中提取出来，然后输入到数据仓库中成为企业数据的一部分。

数据仓库从多种不同数据源，以与关键词和字段定义相容的共同格式，将数据集中在一起。企业一般必须在一个操作系统上进行经营活动，通常不可能（当然也不建议）在该系统上进行密集型的计算或输入/输出数据挖掘操作。但无论如何，操作系统以某种格式存储数据，这种格式是为最优化操作任务的性能而设计，通常这种格式不太适合像数据挖掘之类的决策支持工作。数据仓库应该是专门为决策支持而创立，以便简化数据挖掘者的工作。

1.4.3 计算能力足以承受

数据挖掘算法通常需要并行处理相当数量的数据，很多也是精深的计算。硬盘、内存、处理器和I/O带宽连续惊人的降价，已经使得曾经昂贵的仅用于政府资助的几个实验室的技术进入普通企业。

主流提供商，如Oracle，Teradata和IBM的并行关系数据库管理软件的成功引入，已经将并行处理能力带入很多公司的数据中心。这些并行数据库服务器平台为大规模的数据挖掘

提供了极好的运行环境。

1.4.4 客户关系管理的兴趣增强

在各行各业，许多公司已经开始认识到客户对业务非常重要，客户信息是他们的宝贵财富之一。

1. 每种业务都是服务业务

对从事服务业的公司来说，信息意味着竞争优势。这就是为什么连锁旅馆记录你对不吸烟房间的倾向，汽车租赁公司记录你偏好的车型。另外，传统理解认为自己不是服务提供商的公司也开始有另外的思考。汽车经销商出售的是汽车还是运输能力？如果是后者，当你的汽车在修理时，经销商就应该提供一辆代用车（现在很多都如此）。

甚至日用品也可以增加服务内容。家庭燃油供应公司监控你的用油状况，当你需要更多油时他们会递送，与那些期望你记住在油箱用干或导管封住之前打电话安排递送的公司相比，会售出更多产品。信用卡公司、长途服务提供商、航空公司和所有产品种类的零售商人，经常在服务和价格两个方面进行竞争。

2. 信息是产品

很多公司发现，他们拥有的关于客户的某些信息不仅对自己非常有用，对别人也同样有用。在有忠诚卡计划的超级市场中，也有消费者货物包装产业喜欢知道的信息，即关于谁购买了哪些产品的信息。信用卡公司也有航空公司想知道的信息，即谁购买了大量的机票。超级市场和信用卡公司都处在信息经纪人（中间人）的位置。当超级市场承诺可以给合适的购物者更高兑换率时，超级市场可以收取消费者货物包装公司的更多费用来打印优惠券。信用卡公司可以要求航空公司针对经常乘飞机的人们进行促销，吸引以前乘坐其他航空公司飞机的人。

Google 知道人们希望在网络中寻找什么，这得益于出售赞助链接所获得的信息：保险公司支付费用确保搜索“汽车保险”的人链接到公司的网址；金融服务支付赞助链接费，以便有人搜索短语“抵押贷款”时可以出现该赞助商的链接。

事实上，任何公司在收集有价值的数据时，就处于信息经纪人的位置。Cedar Rapids Gazette 报得益于其在东爱荷华州 22 县的优势位置，为地方商业活动直接提供市场服务，该报利用讣告版面和婚庆告示维持它现有的市场数据库。

1.4.5 商业数据挖掘软件产品已经易于使用

从新的算法首次出现在学术杂志和令人兴奋的会议，到使用这些算法的商业软件变为可用，总有一个时间延迟。从第一个可用产品到其被普遍接受，还有另一个时间延迟。对数据挖掘来说，目前已经到了普遍可用和普遍接受的阶段。

本书讨论的很多技术最初出现在统计、人工智能和机器学习领域。经过大学和政府实验室的几年研究之后，一种新技术开始被商业部门的一些早期接纳者使用。在新技术的这一发展时期，软件一般是以源代码的形式出现，你可以通过 FTP 找到并编译，通过阅读作者的博士论文领会如何使用它。只有一些先驱者采用新的技术成功之后，才开始出现带有使用手册和在线帮助的真正产品。

现在，很多新技术正在开发中，然而扩展和完善已存在的技术更需要投入大量工作。本

书讨论的各种技术均已用于商业软件产品，尽管没有一个单一的产品能够包含所有这些技术。

1.5 目前如何使用数据挖掘

展示这些重要的数据挖掘案例旨在说明本书讨论的数据挖掘技术的广泛应用。这些简介的目的是传达本领域的一些令人兴奋的发展和一些在自己的工作中有益地使用数据挖掘的可能方法。

1.5.1 超级市场成为信息经纪人

销售点扫描器记录了顾客购买的每件货物，忠诚卡计划把这些购买与个人用户相关联，超级市场如今拥有很多客户信息。

Safeway是美国首先开始利用这项技术成为信息经纪人的连锁超市之一。Safeway直接从顾客那里购买住址和人口统计数据，作为回报，客户购物时使用他的忠诚卡得到一定折扣。为获得忠诚卡，购物者自愿披露可以被利用的那些个人信息。

从那时起，购物者每次出示忠诚卡，其交易历史就在某处的数据仓库中被更新。每去商店一次，购物者就提供给销售商更多关于他们的情况。超级市场本身可能对群体模式比对客户的个人行为更感兴趣，如哪些品种一起销售更好，哪些应该放在同一个货架上。摆放在商店过道的产品的制造商对收集到的关于个体的信息最感兴趣。

当然商店向顾客保证为收集的这些信息保密，事实也确实如此。连锁超市不是把经常购买百事可乐的用户名录卖给可口可乐公司，而是根据连锁店所理解的客户购买习惯和客户提供的数据，向某种特定产品的供应商出售可能的客户接触途径。对每个客户名字，Safeway向供应商收取几分钱，以使供应商的优惠券或特殊促销恰好能到达合适的客户手中。因为优惠券兑换也是购物者交易历史文件的一条记录，目标群的精确响应速度也是某种记录。另外，一个特定用户响应某个优惠或者不予回应，都成为未来预测模型的输入数据。

同样地，American Express公司和其他支付卡提供商也出售在账单信封内或信封上的广告空间。他们对广告空间所收费用的高低，直接与他们正确识别用户可能对广告有回应的能力有关，这也正是数据挖掘的实用之处。

1.5.2 基于推荐的商业

英国的Virgin Wines酒店通过其网站www.virginwines.com直接向消费者出售葡萄酒。当新客户首次访问站点时，需要完成一份调查表“葡萄酒向导”，要求每个客户评价不同类型的葡萄酒。这种分级用于创建用户口味的简档。在简档创建期间，葡萄酒向导尝试着推荐不同的产品，消费者有机会选择是或否来细化简档。当葡萄酒向导结束时，站点已经知道消费者的足够信息，可以开始向其推荐消费了。

随着时间的推移，站点追踪客户实际购买的物品，使用这些信息来更新其简档，消费者可以在任何时间重新使用葡萄酒向导更新其简档。他们也可以通过点击“我的酒窖”浏览过去的购买情况。消费者曾经购买或者在站点上评价的任何一种酒都会出现在酒窖中。消费者可以在任何时间评价或重新品评过去购买的东西，提供更多反馈给推荐系统。通过这些推荐，站点可以为消费者提供他们喜欢的新葡萄酒。这给出了像Wine Cask这样的商店建立忠

诚客户关系的有效方式。

1.5.3 交叉销售

USAA 是一个市场定位面向现役和退役军人及其家庭的保险公司，致力于基于信息的营销，包括对普通客户拥有产品数量的两倍使用数据挖掘技术。USAA 拥有客户详细信息记录，使用数据挖掘预测他们处于生命周期的哪个阶段，很可能需要哪些产品。

另一个使用数据挖掘改进交叉销售（cross-selling）能力的公司是 Fidelity 投资公司。Fidelity 维护着一个拥有所有零售客户信息的数据仓库。这些信息用于构造数据挖掘模型，预测另外哪些 Fidelity 的产品可能使消费者感兴趣。当一位客户打电话给公司时，销售代表的屏幕正确地显示出该向哪里引导谈话。

除了改进公司的交叉销售能力之外，Fidelity 的零售市场数据仓库还允许金融服务职能部门建立模型，研究究竟是什么造就忠实的消费者，进而增强消费者的持久力。这些模型曾经一度促成 Fidelity 保留支付服务，否则该服务可能早已被取消。与普通消费者相比，使用该项服务的人更不容易把他们的业务转向其他竞争对手，而取消这项服务意味着把有利可图的忠实消费者群体推到其他公司。

客户关系管理的中心原则是，与市场份额相比，关注“钱包份额”或者“消费者份额”（即每位消费者的业务数量）可以获得更多利润。从金融服务业到重工业，很多创新型的公司正在使用数据挖掘来增加每位消费者的价值。

1.5.4 抓住好的客户

客户可以用极小的代价自由改变供应商，竞争对手渴望引诱客户到自己的公司，所以任何行业都需要利用数据挖掘提升客户的持久性。银行称为内耗，无线电话公司称为搅局。无论称为什么，这都是每一个企业面临的重要问题。通过了解谁可能离开和为什么离开，针对好的客户提出合适的手段，可以形成保证客户持久性的方案。

在成熟的市场，引入一个新客户往往比保持一个现有客户代价更高。然而用于保持消费者的动机往往代价昂贵。哪些消费者应该得到该激励？哪些消费者即使不需要这个诱因也可保持？应该允许哪些消费者离开？数据挖掘是断定这些情况的关键。

1.5.5 淘汰差的客户

在很多行业，为有些客户的付出高于客户的回报，这可能是那些耗费了大量的客户支持资源而购买量不足的客户，或者是一些持有信用卡而极少使用的讨厌的家伙，当他们购买时，当然会付清全部款项，但银行仍然必须每月给他们邮寄对账单。更糟的情况，他们可能是那些已申请破产却欠你很多钱的人。

用于发现最有价值顾客的数据挖掘技术，同样也适于发现应该拒绝哪些人的贷款，哪些人应该等待最长时间，哪些人应该总是被安排在靠近发动机的中间座位（或者这只是我们的偏执状况？）。

1.5.6 变革一个行业

1988 年，信用卡发行者拥有的客户信息是他们最宝贵的财富，这一论点被称为是革命

性的创意。Richard Fairbank 和 Nigel Morris 游说了 25 家银行，直到 Signet 银行答应试试这个创意。

Signet 通过各种渠道收集行为数据，用于创建预测模型。利用这些模型，开办了非常成功的转账系统，改变了信用卡行业的工作方式。Signet 在 1994 年推出了 Capital One 银行卡业务，目前已跻身为前十位信用卡发行商。Signet 大胆使用数据挖掘技术加速了公司的成长，同时也保证了 Capital One 在本行业中最低的贷款损失率。目前数据挖掘技术已经成为所有主要信用卡发行商的市场策略的核心内容。

信用卡部门已经率先在银行收费中使用数据挖掘，其他部门也不落后。Wachovia 是一个总部设在北卡罗莱纳州的大银行，它已经将数据挖掘技术用于预测哪些客户会在近期搬家。对大多数人来说，搬到另外一个镇上的新家后，一般会关闭旧的银行账户而选择另外的银行新开一个账户。Wachovia 推出了一些举措来改善客户的持久性：找出要搬家的客户，然后让他们很容易地把业务转到当地的 Wachovia 支行。通过这样的措施，不仅客户持久度有了显著提高，还开发出了利润可观的再分配业务：除了建立新的银行账户以外，Wachovia 现在还在新的支行办理代缴煤气费、电费以及其他相关服务。

1.5.7 其他

这些应用可以让你感觉利用数据挖掘技术可以完成哪些工作，但这尚未包罗所有可能的应用。本书中讲到的数据挖掘技术已经被用于发现类星体、设计军服、查出号称“超纯”的伪劣二次压榨橄榄油、教机器大声朗读以及识别手写字等。毫无疑问，它们在本世纪中还会继续被用于处理那些正在成长或繁荣的业务。在下一章中，我们会讨论在商业活动中如何通过数据挖掘的良性循环来应用数据挖掘技术。

1.6 小结

数据挖掘是客户关系管理系统分析的重要组成部分之一。客户关系管理系统分析的目标是最大程度地再创造这种紧密的学习关系，使得运转良好的小企业与客户合作愉快。公司与客户的交流会产生大量的数据，这些数据一般是由交易处理系统（如自动售货机、电话交换机记录以及超市的扫描器文件）收集而来，然后将这些数据集中、清理、汇总后进入客户数据仓库。设计良好的客户数据仓库包含客户与公司交流的历史记录，成为公司存储内容的一部分。将数据挖掘工具用于处理这些历史记录，可以帮助公司将来更好地服务于客户。本章给出了几个使用数据挖掘的商业案例，诸如优化优惠券设计、推荐销售、交叉销售、客户保持以及降低信用风险等。

数据挖掘本身就是从大量数据中发现有用模式和规则的过程。本章中引入并定义了 6 个常见的数据挖掘任务：分类、估计、预测、关联分组、聚类和建立简档。本书的其余部分分析了完成这 6 类任务的许多数据挖掘算法和技术。这些技术必须成为大型商业活动的有机组成后，数据挖掘才能成功进行。这种一体化的过程就是下一章的主要内容。

第2章 数据挖掘的良性循环

19世纪初，纺织厂成为工业革命的成功题材。为了利用水力，英格兰和新英格兰发展中城镇和都市的纺织厂都沿河而建。奔腾的河水驱动架在河上的水轮，从而驱动着纺纱、针织和编织机器。在长达一个世纪的时间里，水力驱动的纺织机械一直是工业革命的标志。

现在，商业界已发生巨变，老厂区已经成为历史古迹，沿河而建的老厂房变成了仓库、大型购物中心、艺术馆和计算机公司。就连制造公司，在服务行业创造的价值也时常超过货物本身的价值。一家著名的国际水泥制造商（Cemex公司）的广告，给我们留下了很深的印象。其创意是：把混凝土当作服务。那则广告不是集中宣传水泥的质量、价格和实用性，而是在河上画了一座桥，展示水泥是服务的理念：人与人之间通过“水泥”建成的桥梁彼此沟通。混凝土能当作服务吗？这可是一个非常新颖的想法！

使用电力或机械力不再是成功的标准。对于大规模的产品销售，客户交互数据就像是新型水力资源。由于服务业和制造业之间的界线正变得模糊不清，知识除了驱动制造业经济的涡轮，也驱动着服务业经济的涡轮。从数据处理得到的信息，用于对客户进行划分可帮助市场营销；用于改进产品设计可以满足客户的实际需求；用于了解和预测客户的倾向，可以改善资源配置。

数据被看做是大多数公司核心业务处理的中心内容。无论是哪一个行业（比如零售、电信、制造、公共服务、运输、保险、信用卡和银行业等）的操作系统中，任何交易都生成数据。各种外部来源使系统内数据大量增加，这些外部来源包括零售客户的人口统计学数据、生活方式、信用信息，以及企业客户的信用、财务、交易信息。数据挖掘的目标就是，在这些数以百万亿计的字节中，发现潜在有价值的模式（pattern）。但是，仅仅找到模式还不够。你必须对这些模式做出反应，对它们进行处理，最终将数据转化为信息，将信息转化为行动，最后将行动转化为价值。简而言之，这就是数据挖掘的良性循环。

为实现数据挖掘的这个目标，数据挖掘需要成为一个实质性的业务过程，并且融入到市场调查、销售、客户支持、产品设计和库存控制等其他过程。良性循环就是使数据挖掘更深地植根于业务的环节之中，将焦点由探索机制转到以发现为基础的行动上。纵览本章和全书，我们将从数据挖掘的探讨中获得可操作的（actionable）结果。

市场营销文献使得数据挖掘显得如此容易，即，只要运用学术界精英创造的自动化算法，如神经网络（neural network）、决策树和遗传算法（genetic algorithm），你就可以走向无数的成功。尽管算法重要，但数据挖掘的解决方案并不仅仅是一系列有效的技术和数据结构。各种技术必须用于合适的领域，作用于正确的数据。数据挖掘的良性循环是一个反复学习的迭代过程，该过程以上次结果为依据，随着时间的推移而完善。成功地运用数据，可以使一个企业由被动反应转变为先发制人，这就是数据挖掘的良性循环（virtuous cycle）的作用。本书作者就是应用这些技术去争取最大收益的，后面将进一步描述这些技术。

本章从描述一个应用数据挖掘技术的简短案例开始，一直到解决现实商务问题的案例结束。利用案例研究来介绍数据挖掘的良性循环。数据挖掘表现为商业活动过程中一个持续的过程，一个数据挖掘项目的结果变成下一个项目的输入。每个数据挖掘项目都经过四个主要

阶段，它们一起构成了良性循环的完整过程。在介绍完这些阶段之后，将采用案例研究进一步阐述。

2.1 商业数据挖掘案例研究

从前，一家银行存在一个业务难题，他们的特别的商业产品——家庭抵押贷款额度，不能吸引好的客户。为解决这个问题，银行可以采取几种办法。

比如，银行可以降低贷款利率。如果这样，银行能吸引更多的客户，但是以降低利润为代价来增加市场份额。如果降低贷款利率，现有客户也可能转向更低贷款利率，进一步降低了这家银行的利润。更糟糕的是，假若开始时确定的利率具有合理的竞争性，降低利率可能会招致恶意的客户——不忠实客户。竞争对手略施小恩小惠，便可轻易将他们收买。下面"赚钱或者赔钱"部分中所谈到的就是挽留忠实客户的问题。

这个例子中，在经历了几次直接邮寄活动所产生的令人失望的结果后，美国银行急于扩张家庭抵押贷款的业务量。美国消费者资产协会（NCAG）决定采用数据挖掘来解决这个问题，这是引入数据挖掘良性循环的一个很好的案例。（在此要感谢 Larry Scroggins 先生允许我们使用他撰写的《美国银行案例分析》一书的部分资料。我们也从与 Hyperparallel 资料分析公司的 Bob Flynn、Lounette Dyer 和 Jerry Modes 的谈话中受益匪浅。）

2.1.1 识别商务挑战

美国银行需要向客户做好家庭抵押贷款的宣传工作。根据一般常识和商业顾问的意见，他们达成以下共识：

- 有孩子上大学的家长，想通过家庭抵押贷款借款支付学费。
- 高收入但收入不稳定的人，想通过家庭抵押贷款使其收入削峰填谷。

赚钱或者赔钱？

家庭抵押贷款产品的利率可以给银行带来收益，但是有时公司要与亏损的服务项目作斗争。例如，Fidelity 投资公司曾打算将支付服务系统进行拍卖，因为该服务系统一直处于亏损状态。但调查证实，Fidelity 投资公司多数忠实的、有利可赚的客户在使用支付服务系统，最终的分析挽救了服务系统。虽然支付服务系统亏损，Fidelity 投资公司却凭借这些客户的其他账户挣了许多钱。客户毕竟信任他们的金融机构，让他们帮助支付账单，该机构对这类客户有很高的信用度。

削减这样的增值服务项目，导致最好的客户到别处寻找较好的服务，这样做无意之中降低了公司的收益。

家庭抵押贷款产品的市场营销文献反映了潜在客户的观点，电信市场的客户列表清单也证实了类似观点。这些观点解释了前面提到的那些令人失望的结果。

2.1.2 应用数据挖掘

美国银行与来自 Hyperparallel 公司（一个数据挖掘工具供应商，后来被 Yahoo 公司收购）的数据挖掘顾问一起工作，决定用一系列数据挖掘技术来解决这个问题。他们不缺乏数据，多少年来，美国银行一直将其数百万的零售客户数据存储在一个巨大的关系数据库中，

这个数据库安装在美国 NCR/Teradata 公司生产的大型并行计算机上。由 42 个记录系统提供的数据被筛选、转换（transform）、调整后，馈入到公司数据仓库。美国银行利用这个系统，能参透与银行保持联系的每位客户的所有关系。

这个历史数据库确实名副其实，数据库中的有些记录可以追溯到 1914 年！近年来的客户记录大约有 250 个字段，除银行内部数据外，还包括人口统计信息字段，例如收入、子女数量和家庭类型。客户的这些属性汇集成客户独一无二的特征，然后采用 Hyperparallel 公司的数据挖掘工具进行分析。

决策树导出划分现有银行客户的规则，把客户分为两类，即可能或不可能对提供家庭抵押贷款做出反应。经反复检验数以千计购买产品和数以千计没有购买产品的客户数据，决策树最终获得判定不同类客户之间差别的规则。一旦发现这些规则，利用得到的模型可以给每个潜在客户记录增加另一个属性。这个属性即好的潜在客户标志，就是由数据挖掘模型生成的。

下一步使用后续的模式查找工具，可以确定客户什么时候最有可能需要这种贷款。这种分析的目标就是发现过去曾经频繁处理成功诱因的一系列事件。

最后，应用聚类工具自动将具有相似属性的客户分成不同组。在某一点上，这个工具发现了 14 个客户簇，其中许多簇似乎没有特别的兴趣。但是有一个簇的兴趣十分浓厚。这个簇具有两个十分令人费解的特点：

- 这个簇中 39%的人同时拥有企业和个人账户。
- 根据决策树分类，这个簇中的客户占到了家庭抵押贷款可能响应者的四分之一。

这些数据提示好奇的数据挖掘者，上述簇中的客户有可能使用家庭抵押贷款来从事商业活动。

2.1.3 按照结果采取行动

利用这个新的发现，美国消费者资产协会（NCAG）和银行的零售分支机构联合采取下列行动：他们组织市场调查，与客户面谈。现在，银行又增加了一个想弄明白的问题，“贷款收入将被用于从事商业活动吗?”市场调查的结果证实了这个由数据挖掘引出的问题。因此，美国消费者资产协会打消了顾虑，继续瞄准他们的家庭抵押贷款营销活动。

顺便指出，市场调查和数据挖掘时常被应用于类似的目标——对客户获得更多的了解。尽管市场调查非常有成效，但是也存在一些缺点：

- 响应者（responder）不可能代表全部人口，也就是说，同一组响应者会有偏差，尤其是在过去营销工作比较集中的地区，这样就会形成所谓的*机会样本*（opportunistic sample）。
- 客户，特别是不满意的客户和以前的客户，没有理由帮助市场调查，或者诚实对待市场调查。
- 对任何给定的行为，可能是由多方面原因引起的。例如，银行办事处关闭，客户支票被退回，以及在 ATM 机前长时间等候等原因，都有可能使客户放弃那家银行。尽管时序可能更重要，但是市场调查也许仅仅了解了其中的大概原因。

尽管存在这些缺点，与现有客户和以前的客户进行面谈，也可以深入了解其他任何方式都无法得到的情况。美国银行的这个例子说明，这两种方法可以彼此互补，和谐一致。

提示：当对现有客户进行市场调查时，运用数据挖掘技术，把已经掌握的客户信息考虑在内是一个不错的主意。

2.1.4 测试效果

美国银行发现，由于新近采取了一些活动，家庭抵押贷款的响应率从0.7%上升到7%。按照集团副总经理Dave McDonald的说法，数据挖掘结果表明，在零售业务方面银行缺少的是从大规模营销机构向学习型机构的转变。“我们希望能达到这样一个目标：坚持不懈地执行市场营销计划——不仅仅是每季度发邮件，同时相应地推出多个计划。”他设计了一个营销过程的循环图：在图中，运行的数据被传送到一个快速分析处理的过程，然后产生执行和测试计划，这些计划又会产生使营销过程更加完善的另外一些数据。简而言之，这就是数据挖掘的良性循环。

2.2 何谓良性循环

美国银行的例子展示了数据挖掘的良性循环实践，图2-1显示了循环的四个步骤：

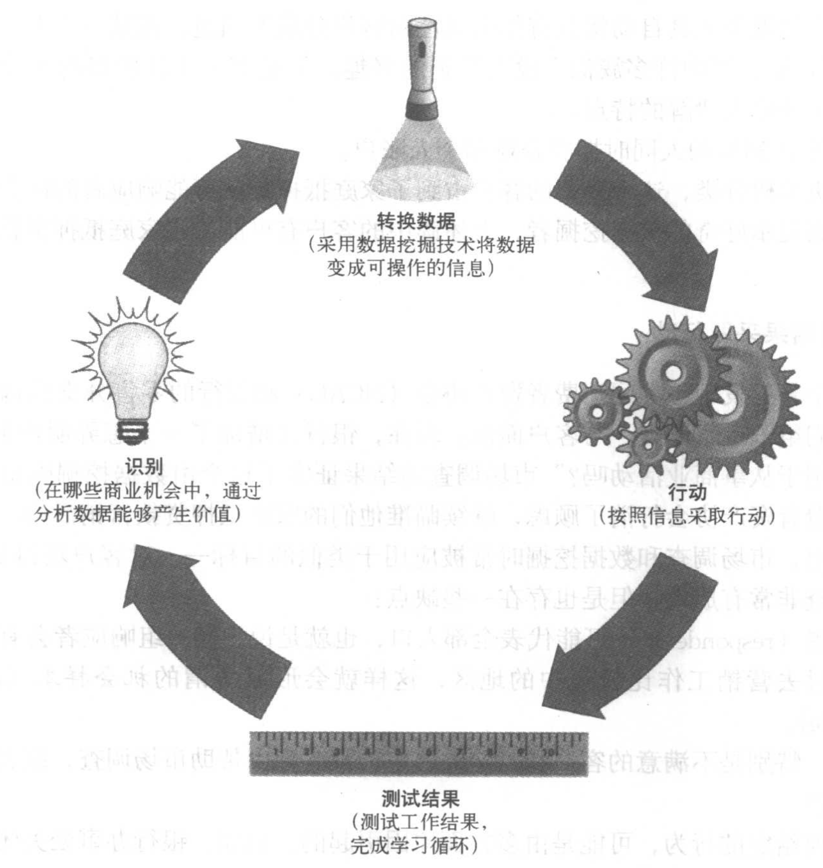


图2-1 数据挖掘的良性循环注重商业结果，而不仅仅是利用先进的技术

第一步，识别商务问题。

第二步，应用数据挖掘将数据变成可操作的信息。

第三步，按照信息采取行动。

第四步，测试结果。

上述步骤说明，成功的关键是将数据挖掘融入到商业过程，并且鼓励数据挖掘人员和使用结果的商业用户密切配合和沟通。

2.2.1 识别商业机会

数据挖掘的良性循环从识别正确的商业机会开始。不幸的是，很多好的统计师和能力强的分析师所做的工作，实质上是浪费时间和资源，因为他们手头的工作对商业活动并没有帮助。优秀的数据挖掘人员要避免出现这种情况。

主动注重实效，将会避免分析工作的浪费。很多普通的商业过程是数据挖掘的很好题材。

- 规划新产品介绍
- 策划直接营销活动
- 了解客户流失行为
- 评估市场营销试验的结果

通过瞄准不同群体，调整信息等手段，让企业经理做出比较有水平的决定，这都是数据挖掘提升现有商务活动业绩的例子。

为了避免分析尝试的浪费，测试采取任何行为所造成的影响也是很重要的，这可以判断数据挖掘工作本身的价值。如果我们不能测试数据挖掘的结果，那么就无法从工作结果中获得经验，也就没有良性循环可言。

对商业活动中过去的尝试等工作进行测试，也会发现数据挖掘的机会：

- 什么类型的客户对上次活动做出反应?
- 最好的客户在哪儿?
- 在自动取款机前长时间等待，是导致客户流失的原因吗?
- 好的客户群体使用客户支持系统吗?
- 应该与 Clorox 公司生产的漂白剂一起捆绑推销什么产品?

开始数据挖掘工作的另一个好的办法就是会见商务专家。从事商业活动的人们可能不熟悉数据挖掘，他们也可能不懂得如何按照数据挖掘结果采取行动。这种会面，既是专家向企业解释数据挖掘价值的过程，也可为双向交流提供平台。

我们曾经参加过一家电信公司的系列会面活动，目的是讨论分析呼叫详细记录（每位客户已经呼叫的记录）的价值。在其中的一次会面中，与会者不能理解这些事情到底有什么用处。后来，一位同行指出，呼叫数据中隐含着客户在家使用传真机的信息（具体细节讨论参见第10章链接分析）。滴答！使用传真机是人们是否在家工作的一个很好线索。并且利用这个信息，可以设计一个针对在家工作群体的产品包。没有我们的提示，这类销售企业决不会考虑通过研究数据，发现这个重要的信息。将技术和商业结合起来，使极有价值的商机突现出来。

提示： *当和商业用户谈论数据挖掘机会时，一定要记住，他们关注的是商业问题而不是技术和算法。要让技术专家关注技术问题，商务专家关注商业问题。*

2.2.2 挖掘数据

数据挖掘，即本书的核心所在，就是将数据转化成可操作的结果。成功的数据挖掘是让数

据有商业价值，而不是运用特别算法或者工具。大量的陷阱干扰着数据挖掘结果的应用能力：

- 坏的数据格式，例如，在结果中客户的地址不包含邮政编码
- 混淆数据字段，例如，在一个系统中，发送日期的本意是“计划发送日期”，而在另一个系统中却是“实际发送日期”
- 缺乏功能性，例如，呼叫中心的申请表不允许有个性化的注解
- 法律分歧，例如，当放弃贷款时，必须提供法律依据（并且“我的神经网络告诉我就是这样”是不被接受的，要讲法律依据）
- 机构因素，因为有些商业集团不希望改变他们的运作方式，特别是没有动力的时候
- 缺乏时效性，因为结果出来的太晚，不具有可操作性

正如图 2-2 所示，数据来源多样，有多种格式，出自若干系统。找出合适的数据源，将它们汇总在一起，这是数据挖掘成功的关键。每一个数据挖掘项目都有数据问题：不一致的系统、表格的关键字与数据库不匹配、间隔几个月记录会被重写等。对数据的抱怨往往是无法做任何事情的第一借口。真正的问题应该是“利用已有的数据能干什么？”这是本书后面讨论的算法会讲到的问题。

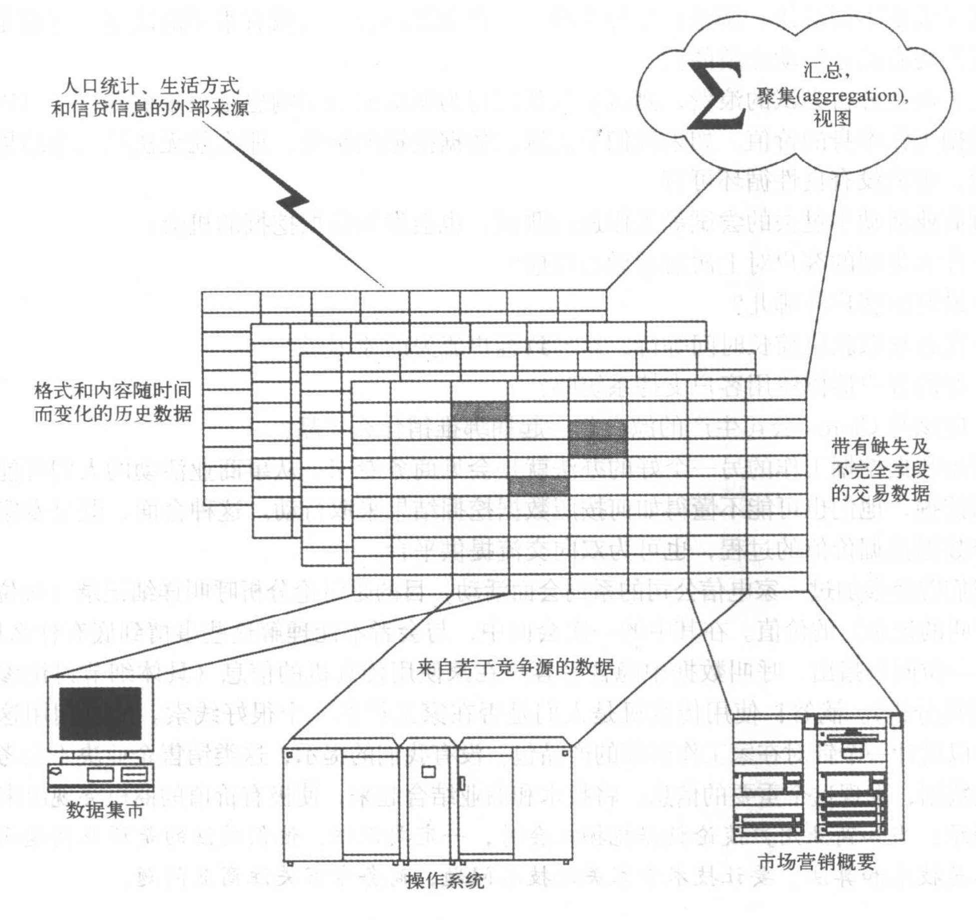


图 2-2 数据从来都不完整清晰。数据以各种形式存在，来自很多内部和外部的数据源

一个无线通信公司在获得一台大型服务器和数据挖掘软件包后，曾想整合资源，成立了

数据挖掘组。在后期，他们联系了数据挖掘人员来帮助公司调查数据挖掘的机会。在这个过程中，我们发现，客户流失的一个关键因素是过度通话：在第一个月期间，新客户打电话太多。客户在拿到第一份话费单时，有时是下一个月的中间，才得知通话过度。客户那时已经积累了大笔话费，十分不高兴。不幸的是，客户服务机构也必须等到同样的清算周期时，才发现客户过度通话。也就是说，没有时间提前来应对这种局面。

然而，初期的数据挖掘组有很多资源，并且已经识别出合适的数据来源，利用某些相对简单的程序运算，在刚刚出现过度通话的几天内是有可能识别这些客户的。利用这个信息，客户服务中心能联系到正处于过度通话风险的客户，并且在出第一张话费单前，将这些客户转移到合适的话费套餐。这个简单的系统是数据挖掘的一个很大胜利，仅仅因为有了一个技术水平高、装备软硬件和具有访问权的数据挖掘组，他们装配了这个关键的系统，把可能性变成了现实。

2.2.3 采取行动

采取行动是数据挖掘良性循环的目的。前面已经提到过，行动可以有很多不同的方式。数据挖掘提高了商务决策的水准。随着时间的推移，我们期望，更高水平的决策产生更好的结果。

无论如何，行动通常要与商务活动的安排相一致：

- 通过直接邮寄、电子邮件和电话推销等手段，给客户和潜在的客户发送信息；使用数据挖掘，不同的人群可以得到不同的信息
- 为客户服务划分优先次序
- 调节库存水平，等等

数据挖掘的结果需要馈入到与客户接触和影响客户关系的商业过程中去。

2.2.4 测试结果

在前面，已经强调了测试结果的重要性。尽管它很重要，但在数据挖掘的良性循环中该阶段极有可能被忽略。尽管测试和不断改进的意义被广泛认可，但实际上往往并没受到应有的重视。有多少商业案例得到贯彻执行？实际上没有人回过头来了解现实与计划匹配得到底怎么样。个人可以通过采取“比较和学习、提问为什么计划与现实匹配或不匹配、愿意获知早期设想是个错误”等措施，全面改进自己的做法。对个体起作用的方法对企业同样也起作用。

在识别商务问题的时候，首先必须考虑结果的测试。怎样才能测试结果呢？为激励产品销售，公司开展赠送优惠券活动，毫无疑问要测试优惠券返回率。然而，持优惠券的购买者可能本来无论如何都打算购买该产品。另一种合适的度量方法是在特定商店或地区增加销售量，这种增长就与特定营销工作相挂钩。由于这些方法需要大量详尽的销售信息，做出这种测试可能是困难的。然而，如果目标就是要增加销售量，就必须有直接方法测量它，否则的话，营销工作完全可能变得“闹闹哄哄、令人愤怒和毫无价值”。

也许这种市场营销干预了几个月之后，包含概要内容的标准报告才会提交出来。即使这类报告中含有重要信息，销售经理也可能不能从报告中注意到这些重要信息。要理解市场营销活动对客户保持度的影响，就要更长时间地追踪已经采取的市场营销工作的结果。设计优良的联机分析处理（Online Analytic Processing，OLAP）技术（在第 15 章中将专门讨论），

对销售团队和销售分析师会很有帮助。然而，对于一些问题，可能需要有更详细的信息。

把每次数据挖掘的尝试作为小型商业个案来考虑，是一个很好的主意。通过比较预期结果和实际结果，可以为下一个良性循环周期找出可能的机会。我们时常过分忙于处理下一个问题，以至于没有精力测试当前尝试的成功状况，这种做法是错误的。每次数据挖掘尝试，无论成功与否，都会对下一次的尝试提供经验教训。问题是，需要测试什么和如何进行测试，这些结果为将来的应用提供最好的素材。

作为一个示例，让我们首先从测试一项有既定获取目标的市场营销活动开始。规范的测试指标是响应率：既定的活动对象当中多少人有实质性的反应？这种活动会收集到很多的信息。对这种获取目标的市场营销活动来说，未来有使用价值的一些问题如下：

- 该活动会波及到或带来可赚钱客户吗？
- 客户如期望的那样保留住了吗？
- 通过此项活动得出最忠实客户的特征是什么？老客户的人口统计分析档案可以用于潜在客户中。在有些情况下，这种分析应该限定在那些由外部来源提供的特征，以便数据挖掘分析的结果能成为实用的可购买的客户名单。
- 这些客户购买其他产品吗？企业中不同的系统是否能发现一个客户购买多种产品的情况？
- 某些信息或产品收到的效果是否比其他的好？
- 活动所波及的客户对于从其他渠道得到的信息有反应吗？

所有这些测试结果都能为将来的决策提供依据。为了在将来做出更好的决策，通过学习过去的事情，数据挖掘信息把过去和未来行为联系在一起。

一项特别的测试是终生客户价值（lifetime customer value）。顾名思义，是指在整个客户关系过程中，对客户价值的大致估计。有些行业已经开发了十分复杂的模型用于估计终生客户价值，有一些即使没有复杂的模型，也可以进行短期估计（如1个月、6个月和1年以后），这种估计也被证明是非常有用的。客户价值将在第4章详细讨论。

2.3 良性循环环境下的数据挖掘

一家有代表性的美国大型区域电话公司拥有数百万客户。它的总机房中有数百或数千个交换机，这些交换机通常分布在几个州，横跨多个时区。每个交换机能同时处理数千个电话，包括比较先进的功能，如呼叫等待、电话会议、呼叫转移、语音邮件和数字服务。在已经开发的最复杂计算设备中，只有少数制造商能生产这种交换机。由于供应商不同，这种典型的电话公司一般拥有多个版本的不同交换机，其中的每一台交换机对每次通话和通话尝试以自己的格式提供大量数据，其数据量每天达数万兆字节。另外，每个州都有影响该行业的地方规章，更不用说联邦政府还经常调整法律和规章。再有，电话公司向客户提供数千种不同的电话套餐，面向客户从临时用户到财富100强公司不等，这些都大大增强了数据的多样性。

账单处理是维持企业的生计所在，是企业主要收入来源，这家公司或任何类似的大企业该如何管理账单处理？答案很简单：要非常小心！许多公司已经制定了详细的流程来实现操作的标准化，他们有管理的政策和程序。这些流程是强有力的，即使在企业重组、数据库管理员休假、计算机暂时停机、法律法规修改以及交换机升级的时候，账单照样会送到客户手里。假如一个企业能在每个月将账单准确无误地送到数以百万的居民、企业和政府客户手

中，可以肯定地说，该企业将数据挖掘应用到决策过程应该是件非常容易的事情。情况真是这样吗？

大型企业为经营业务，已经积累了几十年开发和执行关键任务应用程序的经验。数据挖掘不同于典型的操作系统（参见表 2-1），运行良好的操作系统所需要的技巧不一定能产生成功的数据挖掘业绩。

表 2-1　数据挖掘系统与典型商务操作系统的区别

典型操作系统	数据挖掘系统
对历史数据的操作和报告	对历史数据的分析，时常作用于最近的数据以决定未来的行为
可预测和周期性的工作流，显著特点是与日历挂钩	不可预测的工作流，取决于商务和市场营销需求
使用有限的、企业范围的数据	（一般来说）数据越多，结果越准确
关注商业因素（例如账户、地区、产品代码、通话时间等），而不是客户	关注可操作的实体，例如产品、客户、销售地区
响应时间通常以秒/毫秒计量（用于交互式系统），而等待报告需要几周或几个月	迭代过程，响应时间通常以分钟或小时计量
数据记录系统	数据复制
可描述性和可重复性	创造性

首先，数据挖掘解决的问题不同于操作问题。*准确地说，数据挖掘系统并不试图复制以前的结果。*事实上，复制以前的做法会带来灾难性的结果，导致市场营销活动的对象总是同一群人。通过分析数据，你并不想看到大量客户与以前市场营销活动所涉及的客户特征相匹配。数据挖掘的过程需要考虑这个问题，而不像操作系统那样一遍一遍复制同样的结果——是否完成通话、发送账单、授权信用购买、跟踪库存，或是其他无休止的日常操作。

*数据挖掘是一个创造过程。*数据具有很多不是没用就是简单描绘当前业务策略的明显相关性。例如，一家大型零售企业的数据分析显示，签订维修合同的人也极有可能购买大件家具。如果零售企业不分析家具连同维修合同销售的有效性，有这种信息比没有更糟糕——待讨论的维修合同只是与大件家具一同出售。花费数百万美元购买硬件、软件和聘请分析师，却发现这种结果，纯粹是浪费资源，这些资源完全可以应用到商业的其他地方。分析师需要知道什么对商业有价值，并且知道为了获得巨大收益如何整理数据。

*数据挖掘结果随时间而变化。*模型渐渐跟不上时代的变化，最终变得毫无价值。原因之一，就是数据迅速老化，市场和客户也瞬息万变。

*数据挖掘向其他可能需要改变的过程提供反馈。*商界做出的决策，时常影响当前的过程以及与客户的互动。通常情况下，观察数据会发现操作系统的瑕疵。修正这些瑕疵可以增进对未来客户的了解。

本章其余部分再举出一些实用的数据挖掘良性循环的例子。

2.4　移动通信公司建立恰当的联系

无线通信行业竞争非常激烈。无线通信公司一直尝试采用新办法，从竞争对手那里挖走客户，并培养自己客户的忠诚度。服务的基本内容就是提供物美价廉的产品，因此无线通信公司考虑吸引新客户的奇异办法。

本案例所讲的是，一家移动通信公司采用数据挖掘，增强开发客户的能力，希望将客户吸引到公司新的服务项目。（我们十分感激 Apower Solutions 公司的 Alan Parker 先生提供该案例的很多细节材料。）

2.4.1 机会

这家公司原来想测试一个新产品的市场前景。由于技术原因，他们测试产品的最初覆盖面时只选择了几百个订户——仅占目标客户群的一小部分。

因此，最初的问题是推算谁有可能对这种新产品感兴趣。这是数据挖掘的典型应用：采用最划算的方法，实现能波及到的响应者理想数量。按照假定，定向市场营销的固定成本看成是不变的，每次联系的支出也差不多是固定值，那么要减少活动的总成本，就必须降低联系的数量。

为确保实验的有效性，公司需要有一定数量的人签约。对于新产品的宣传活动，公司以前的经验是，大约 2%～3% 的现有客户可能做出满意的响应。因此，为达到 500 名响应者的目标，可能需要联络 16 000～25 000 名潜在客户。

如何选择目标？如果给每位预期客户打分，这件事会变得非常容易。假定分值范围为 1 到 100，1 代表非常有可能购买产品，而 100 代表没有可能购买产品。然后，根据得分情况将候选人进行排序，市场营销人员可以顺着这个名单往下数，直至达到想要的响应者数量。如图 2-3 累积增益图所示，联系最有可能响应的人，以较低的联系量，获得期望的响应数量，因此降低了成本。

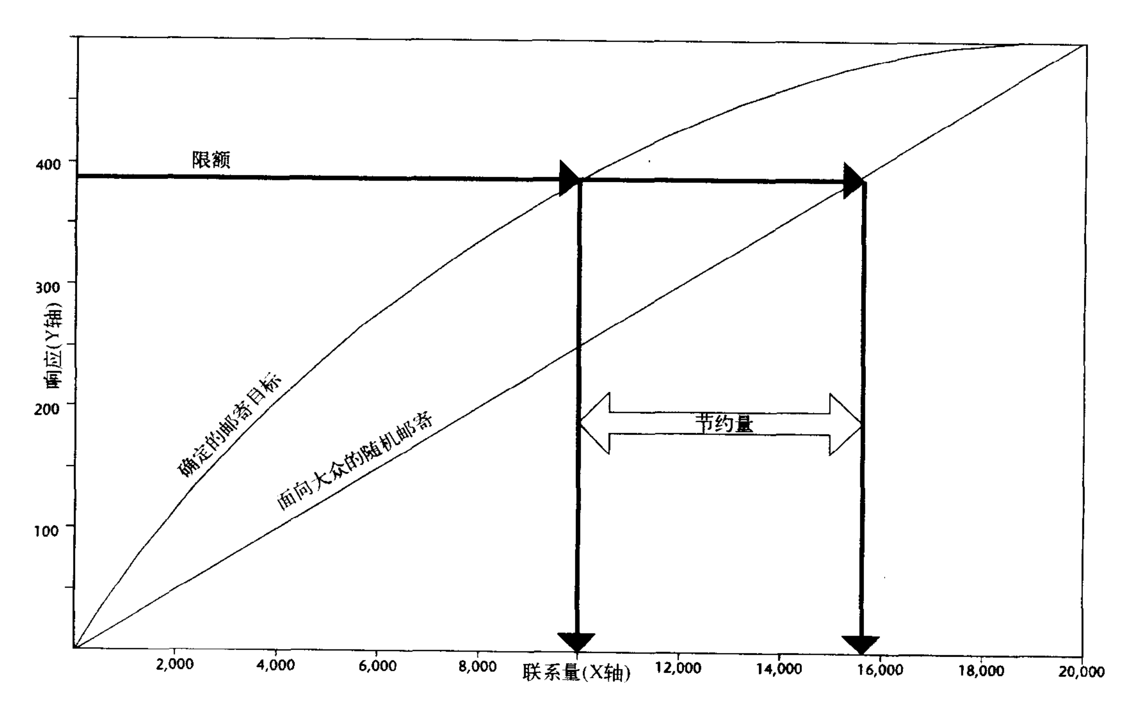


图 2-3 分级潜在客户，利用响应模型，通过确定少数客户目标并且得到同样数量的响应者而达到节约成本的目的

累积增益图将在下一章详细地解释。目前只要知道该曲线是通过把已打分的潜在客户进行排序而得到的就可以了，沿 X 轴方向，靠左边的是最有可能响应的客户，而右边的是最不可能响应的客户。对角线表示的是从所有潜在客户中随机选取样本会出现的结果。从该图可以看出，按照好的响应分值排序，通过接触更少的潜在客户，可以降低定向市场营销活动的成本。

移动电话公司是如何得到这个分值的？当然是靠数据挖掘！

2.4.2 如何应用数据挖掘

多数数据挖掘方法是通过样本获得的，神经网络、决策树生成元或其他方法均来自成千上万的训练样本。每一个训练样本明显标注为响应者、非响应者。在观察足够的类似样本之后，运用工具算出以计算机程序形式表示的模型，然后读取尚未分类的记录，更新响应得分情况或分类。

在本案例中，有待解决的问题是新产品介绍，所以没有已经做出响应的训练样本集合。一种可能的解决办法是，基于对过去任何服务曾做出响应的客户来构建模型。这样的模型能区分拒绝所有电话推销和扔掉所有邮寄宣传品的人，以及那些偶尔对一些服务做出响应的人。这类模型被称为*非响应模型*，对那些真想大范围发送宣传广告的公司会有价值。非赢利、向退休人员提供服务的美国退休人士协会（AARP），应用非响应模型后，节约了数百万美元的邮寄费用。他们以前向至少有一位成员年龄超过 50 岁的所有家庭邮寄广告，现在他们放弃最没有希望的 10%，仍然得到他们想得到的几乎所有响应者。

然而，无线通信公司只是想获得几百名响应者，因此，识别最有希望的前 90% 的模型不可能达到这个目的。相反地，他们借鉴另一个市场上的类似新产品推介，形成训练记录集合。

1. 确定输入

本书描绘的数据挖掘技术中，构建模型过程的核心部分大都自动进行。只要给定一系列输入数据字段和一个目标字段（本案例是指购买新产品），就可以根据输入，找到解释目标的模式和规则。为了使数据挖掘获得成功，必须在输入变量和目标之间建立某种关系。

这实际上意味着，识别、定位和准备输入数据比创建和运行模型经常要花费更多的时间和精力。这是因为，应用数据挖掘工具已经使创建模型变得非常容易。要做好选择输入变量的工作，没有处理商务问题的知识是不可能的。特别是当采用那些声明有能力接受所有数据，并且能自动判断出哪些领域是重要的数据挖掘工具的时候，情况更是如此。行业中有见地的人们所期望的重要信息，往往不能以数据挖掘工具能够识别的方式在原始输入数据中体现出来。

无线电信公司明白选择正确输入数据的重要性。来自几个不同职能部门的专家（包括市场调查、销售和客户支持以及请来的数据挖掘顾问）聚集在一起讨论，寻找可利用现有数据的最佳方法。有三个数据来源可以利用：

- 销售客户信息档案
- 详细的电话呼叫数据库
- 人口统计数据库

目前为止，详细的电话呼叫数据库是三者中最大的一个数据来源，包含目标市场中所有

客户打出和接听电话的每个记录。销售数据库包含简要的客户数据，涉及用法、期限、产品历史、价格方案和付费历史记录。第三个数据库包含关于客户的人口统计和生活方式等数据。

2. 衍生输入字段

通过上述自由讨论和初步分析，在输入到预测模型的客户数据中，增加了几个总结性和描述性的字段：

- 通话时间
- 来电数量
- 呼叫频率
- 影响范围
- 语音邮件用户标识（voice mail user flag）

需要对上述有些字段作一点解释。通话时间（minutes of use，MOU）是衡量客户好坏程度的标准。通话时间越多，越是好客户。公司过去几乎把全部注意力集中到通话时间上，而不顾其他变量因素。但是，通话时间掩盖了很多重要的差异：2个长呼叫好，还是100个短呼叫好？全部主叫好还是半数被叫好？所有呼叫都是同一个号码好还是呼叫很多号码好？上述后面的几个字段将进一步明晰这些问题。

影响范围（sphere of influence，SOI）是另一个重要的度量，它是早期数据挖掘发展出的结果。客户的影响范围，是指在一定时期内，通过电话与该客户进行交流的人数。结果表明：作为群体，影响范围大的客户行为比影响范围小的客户行为在多方面存在较大差异，如呼叫客户服务中心的频率和忠诚度等。

2.4.3 处理行动

把三个来源的数据汇集在一起，用于构建数据挖掘模型。这个模型可以用于识别可能购买新产品的候选人。可以采取两种直接邮寄方式投递：一是按照数据挖掘模型的结果产生名录进行投递，二是投递到采用普通商务方法选出的对照组人群。如图2-4所示，在目标组客户中，有15%的人购买了新产品，而在对照组人群中，只有3%的人购买了产品。

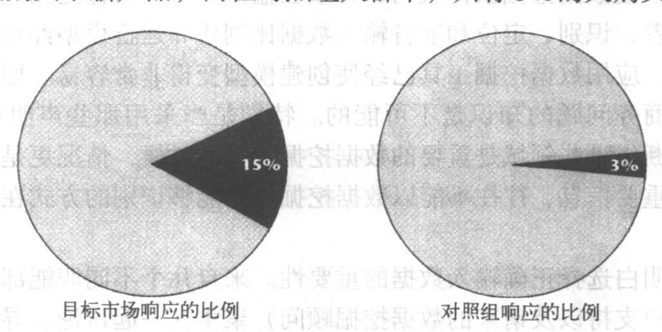


图2-4 这些结果表明数据挖掘的应用非常成功

2.4.4 完成循环

在数据挖掘的帮助下，公司联系到了新产品销售的合适候选人群，但那不应该是故事的

结尾。一旦新的活动结果出来，数据挖掘技术能帮助获得更好的实际反应前景。依靠最初在测验市场上理解到的买主的特征，以及新服务项目开始几个月的使用概况，公司能够在产品推出后的销售产品市场上，更好地寻找潜在客户。

2.5 神经网络和决策树驱动 SUV 的销售

1992 年，在今天可用的任何商业数据挖掘工具面市以前，美国三大汽车制造商之一，要求 Pontikes 管理中心（隶属于南伊利诺斯大学分校，位于美国卡本代尔市）的研究组开发一个“专家系统”，目的是识别可能购买特别的运动型多用途车（sport-utility vehicle，SUV）的客户。（十分感谢 Wei-Xiong Ho 先生，他与南伊利诺斯大学商业管理学院 Joseph Harder 先生一起从事此项工程。）

传统的专家系统是由数百或数千条规则组成的大数据库，这些规则是通过观察和访问擅长特殊任务的人类专家收集来的。在某些特定的领域，例如，在医疗诊断和税收问题方面，专家系统已经获得成功，但是收集规则的难度限制了它们的用途。

为了解决这些问题，南伊利诺斯大学的研究组决定，从历史数据直接生成规则。换言之，他们将用数据挖掘代替专家访问。

2.5.1 最初的挑战

底特律人带给卡本代尔（Carbondale）研究组的最初挑战，就是改善为某个特别车型进行的直接邮寄活动的响应。活动包括向潜在客户发送邀请函，邀请他们参加新车试驾。接受邀请的任何人可以在经销商处免费领到一副太阳镜。问题是很少人将反馈卡寄回或者打免费电话咨询，其中几乎没有人最后确实购买这款车。尽管公司知道，不给那些不响应的人们发送邀请，可以为自己节约很多资金，但他们不知道不响应客户到底是哪些人。

2.5.2 如何应用数据挖掘

正如通常会遇到的那样：待挖掘的数据来自几个不同的信息源，这时第一个挑战就是整合数据，使它们形成一个完整的数据源。

1. 数据

第一个文件“联络文件”（mail file），是一份包含姓名和通讯地址的联系名单。在这份名单上，大约有一百万人，他们都收到过宣传邮件。这份文件含有很少对筛选有益的信息。

在联络文件上，附加了一些邮政编码数据。这些邮政编码可以在商用的 PRIZM 数据库中查到。在这个数据库中，包含了与邮政编码相关联的人口统计学和心理描绘方面的居住区域特征。

另外两个文件含有关于寄回反馈卡或为理解信息拨打免费电话的客户的信息。由于联络文件包含了为每个住址设计的打印在反馈卡上的九字符密钥，所以将反馈卡与原始联络文件联系起来是一件很简单的事情。打电话者则有不止一个问题：打电话者提供的姓名和地址可能与数据库的地址不完全相符；不能保证打电话的人就是联络文件中记录的人，因为收到邮件的人可能已将这种信息转给其他人。

在投放邮件的 1 000 003 人中，32 904 人通过反馈卡做出响应，16 453 人通过拨打免费电话做出响应，初始的总响应率约为 5%。汽车制造商的主要兴趣点当然是那些既对邮件做

出响应，又会购买所推销车型的极少数人。这些资料可以从制造商的销售文件中找到，该文件含有在发出邮件后三个月内所有买主的姓名、地址和购买车型。

采用模糊匹配标准运行自动名称匹配程序发现，在已购车者和收到邮件的人中，大约有22 000对明显匹配。通过手工编辑，把购车者和收到邮件者的交集降为4764人，其中大约有半数的人已经购买了广告推销的车型。图2-5给出了所有数据源之间的比较。

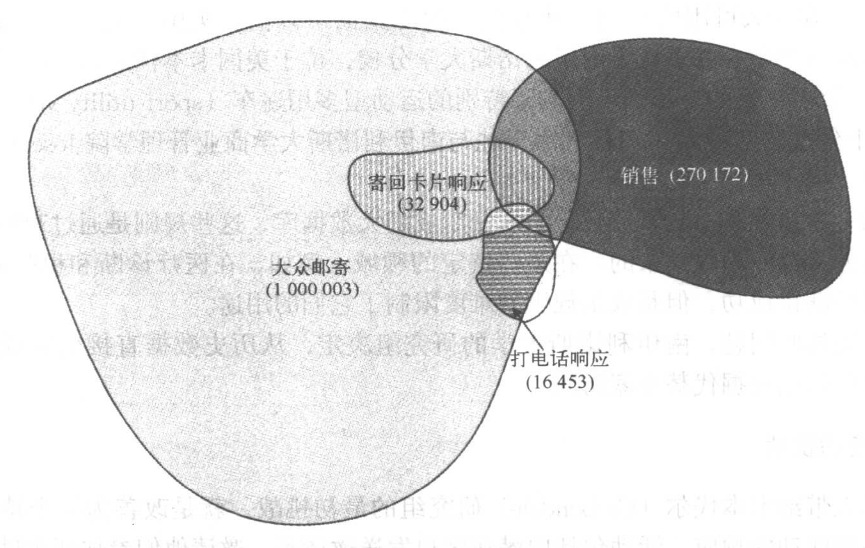


图2-5 在训练集中的潜在客户具有交迭关系

2. 沿标杆向下挖掘

实验设计要求把人群分两个大类：成功和失败。由于设计模糊了有趣的差异，这种严格分类肯定是一项令人质疑的设计。事实上，来到经销商店试一款车，最后却购买了另一种车型的人，应该是不同于非响应者的另一类；已经做出响应但什么车也没买的人，也属于一类。同样，被视为不值得发送邮件却购买了车的潜在客户，更是一个值得关注的群体。

尽管如此，成功的定义是指“收到邮件，并且购买该型号车”，失败的定义则是“收到邮件，但是没有买这款车”。利用决策树和神经网络，我们进行了一系列实验，在多种类型的训练集（training set）中，测试了分析工具。一些训练集返回了数据库中成功的实际比例，而另一些竟然达到10%的成功率。关注程度越高，产生的结果就越好。

神经网络可以对稀少的训练集得到较好的结果，而决策树看来在丰富的训练集上效果更好。研究人员决定把过程分为两个阶段。首先通过神经网络确定谁有可能从公司购买一款车，不计车型；然后应用决策树预测哪类潜在购买者会选择广告推销的车型。两步决策过程被证明是十分成功的。结合神经网络和决策树的数据挖掘模型很少丢掉购买目标车型的客户，同时能够比单独采用神经网络或决策树模型筛选掉更多非购买者。

2.5.3 最终措施

利用能有效波及响应者的模型，公司决定，把减少邮件发送而节约的资金用于增强吸引潜在客户到样车展示室的诱惑。他们向非常小的一个潜在购车群体赠送一双不错的皮靴，而不是向大众赠送太阳镜。结果证明新方法比老办法更行之有效。

2.5.4 完成循环

基于全局的数据挖掘工程显示，即便利用有限的和粗线条的变量以及相当原始的挖掘工具，数据挖掘也能提高定向市场营销活动的效率，即使像汽车这样的大件商品。下一步工作就是收集更多的数据，构建更好的模型，并且反复实验。

2.6 小结

本章从回顾工业革命的驱动力以及英格兰和新英格兰建造大型工厂讲起。这些工厂现在或者已经废弃，或者被推倒，或者改头换面变作他用。水力不再是商业的驱动力，数据已经取而代之。

数据挖掘的良性循环是利用数据作动力，把它转化成可用于商业的结果。就像过去整个工厂的运转曾经是靠水推着轮子转从而驱动机器一样，通过一个组织收集和传播数据也会创造价值。假若把数据类比成水的话，那么数据挖掘就是轮子，良性循环就是将数据动力传播到企业的所有运行过程。

数据挖掘的良性循环是一个基于客户数据的学习过程，起点是识别利用数据挖掘的适当商机。最好的商业机遇就是那些按照数据挖掘指导行动的机遇。如果不采取行动，获取的客户信息其价值就很少或没有价值。

测试行为的结果也非常重要，完成测试也就完成了良性循环的一环，并且通常还会找出进一步做数据挖掘的机会。

第3章 数据挖掘方法论和最佳实践

上一章介绍了数据挖掘作为业务过程的良性循环，讨论中把数据挖掘过程分为4个阶段：

1）识别问题

2）把数据转换为信息

3）采取行动

4）测试结果

从现在开始，应该把数据挖掘作为技术过程来考察了。高层轮廓依然如前所述，但重点转向将商业问题转换成数据挖掘的问题，而不仅是识别商业问题。将数据转换为信息的主题扩大为几个主题，包括假设测试、建立简档和预言性建模。在本章中，采取行动指的是模型部署和评分等技术行为。将一个模型用于指导市场营销行为之前，必须进行实验测量来评价它的稳定性和有效性。

因为全书侧重于方法论，所以本章中介绍的最佳实践活动还会在相关章节给出详细阐述。本章的目的是把方法论集中在一起介绍。

避免中断数据挖掘的良性循环的最好方法是理解其可能失败的情形，然后采取预防性的措施。多年来，作者曾经遇到很多种数据挖掘出现错误的情况。因此，我们发展了一套有效的习惯性方法，即从业务问题的初始描述如何顺利到达能够产生可操作和可测量结果的稳定模型。本章将把这些最佳实践活动总结出的有序步骤，作为数据挖掘方法论来展示。数据挖掘是一个自然的迭代过程，有些过程需要多次重复，但是不应该完全跳过任何一个过程。

数据挖掘的方法越严格就越复杂，如果缺少其中一个步骤，数据挖掘工作就可能失败。本章通过描述各种可能失败的情况，给出了建立方法论需要的内容。下面将首先考虑最简单的数据挖掘方法：使用专门查询来测试假定，然后研究更加复杂的行为，如建立用于评分模型的正规简档、建立真正的预言性模型等。最后，将数据挖掘良性循环的4个步骤转换为数据挖掘方法论的11步。

3.1 为什么需要方法论

数据挖掘是从过去获取知识用于未来更好决策的一种方法。本章介绍的最佳实践方法主要为了避免知识获取过程中出现以下两个不希望的结果：

- 获取不真实的知识。
- 获取真实但无用的知识。

就像水手要学会绕开海上的漩涡和海中的暗礁等危险一样，数据挖掘人员需要了解如何避免这些常见的陷阱。

3.1.1 获取不真实的知识

获取不真实的知识比获取无用的知识更加危险，因为人们可能依据这些不正确的信息做出重要的商业决策。数据挖掘的结果似乎通常是可靠的，因为从表面上看，是基于科学的方

式而获取的。这种可靠性外观很具有欺骗性：因为数据本身可能是不正确的，或者与当前的问题没有关联；发现的模式可能只反映了过去的商业决策，也可能根本什么也没反映；一些数据转换（如汇总）可能破坏或者隐藏了一些重要的信息。下面几节讨论可能导致错误结论的更常见问题。

1. 模式可能不代表任何底层规则

我们经常说数字不会说谎，但是说谎者会乔装打扮。在数据中寻找模式（pattern）时，数据实际上不必撒谎以误导出不真实的结论。有那么多构造模式的方法，因此只要研究足够长的时间，任何数据点的集合都可以揭示一个模式。人们在生活中强烈依赖于不同的模式，即使在不存在模式的时候，我们也努力在寻找它们。当我们抬头看夜空时，看到的不是杂乱无章的星星，而是北斗七星、南十字座或者猎户星座等。甚至有些人看到了用来预测未来的占星术图案或者征兆。广泛接受的各种古怪的协同作用理论是人类需要寻找模式的更进一步的证据。

推测起来，人类变得如此热衷于模式的原因在于，模式通常确实反映了一些现实世界运转的底层原理。月亮的圆缺、四季的更替、日夜的轮转，甚至喜爱的电视节目在一周的某一天的某个固定时间的规则出现都是有用的，因为它是稳定的，因而是有预言性的。可以使用这些模式来决定什么时候种植西红柿是安全的，如何给录像机编好录制节目时间表。另一些模式显然不具有任何预言能力，如果抛一枚硬币一连出现了五次正面向上的情况，第六次抛起仍然有五成的可能会反面朝上。

数据挖掘人员面临的挑战是计算出哪些模式是预言性的，哪些不是。考虑下面这些模式，所有这些都是在一些大众出版物文章中引用的、好像具有预言性价值的模式：

- 非执政党在非大选年的竞选期间获得国会席位居多数。
- 当美国联盟赢得世界职业棒球大赛，共和党人在白宫执政。
- 华盛顿红皮人队赢得最后一个主场比赛，执政党在白宫继续执政。
- 在美国总统竞选中，个子高的人通常会赢。

第一个模式（涉及非大选年）从纯政治的角度看来是可以解释的。因为存在一个潜规则解释，这个模式看起来将会继续，因而有预言性价值。而下面两个包含体育事件的预言，看起来显然没有任何预言性价值。不管共和党人和美国联盟曾经多少次（作者并没有提及这一点）共同分享胜利，也没有理由认为这种关联将会继续。

总统候选人的个子情况又如何呢？至少自1945年Truman竞选成功以来（他虽然个头不高，但是比Dewey要高），Carter击败Ford的那场选举是惟一一位矮个子获胜的选举（只要将“获胜”定义为“获得最多的选票”，2000年大选中身高6′1″的Gore和身高6′0″的Bush竞争仍然适应这一模式）。身高似乎并不应该和总统职位有任何联系，但从另一方面考虑，身高确实与收入和其他社会成功标志有相互关系，因此有意或者无意地，选民会认为个子高的人更适合做总统。正如本章介绍的，正确判断一条规则是否稳定并具有预言性的办法是，比较它在从同一人群中随机选取的多份样本的表现。我们把总统身高的情况作为练习留给读者。通常的情况下收集数据是最难的部分，即使在Google盛行的时代，要收集到18世纪、19世纪以及20世纪落选的总统候选人的身高也并不容易。

发现不能推广的模式的技术术语是*过度适应*（overfitting）。过度适应导致不稳定的模型，这些模型可能某一天起作用，但是另一天却不起作用。建立稳定的模型是数据挖掘方法

论的主要目标。

2. **模型集可能不能反映相关人群的总体状况**

模型集是用于建立数据挖掘模型的历史数据的集合。为使从数据集提取的推论正确，模型集必须反映模型所描述、分类和评分的人群的总体状况。不能正确反映母体数据的样本是有偏差的。使用有偏差的样本作为模型集就会导致获取不完全真实的结果，当然这也是很难避免的。考虑下面的例子：

- 客户不同于潜在客户；
- 市场调查的响应者不同于非响应者；
- 读电子邮件的人不同于不读的人；
- 在网站已经注册的人不同于注册失败的人；
- 公司并购之后，被收购公司的客户未必与并购公司的客户相同；
- 没有缺失值的记录所反映的人群状况，可能不是有缺失记录的人群状况。

现实客户不同于潜在客户，因为他们代表的是在过去的时间里，一直积极响应任何信息、服务和各种吸引客户的促销活动的那些人群。对当前客户的研究可能得到更多同样的结论。如果过去的市场营销活动追逐的是那些市区的富有消费者，那么任何用当前客户与一般人群的比较都可能显示客户应该倾向于富有的城市人。这样的模型可能错过使中等收入的郊区居民成为客户的机会。而使用有偏差样本的结果比仅仅错过营销机会更糟。美国有“经济歧视”的历史，在某些邻近地区有拒绝给予贷款或者保险政策的非法行为。从一个有经济歧视史的公司的历史数据中寻求模式，可能显示某些地区的人们不太可能成为客户。如果未来的销售行为是基于这一发现之上，数据挖掘就会促成这种非法的、不道德行为的永存。

细心关注模型集样本数据的筛选和取样，对成功的数据挖掘至关重要。

3. **数据位于错误的详细层次**

事实说明，在不止一种行业中，在客户要离开前的一个月中商务使用率通常会下降。待我们仔细检查相关数据后，发现这又是一个获取不真实信息的例子。图 3-1 显示的是移动电话用户每月使用分钟数。7 个月来，该用户月平均使用 100 分钟以上。然后在第 8 个月，使用率下降了一半。在第 9 个月根本就没有使用。

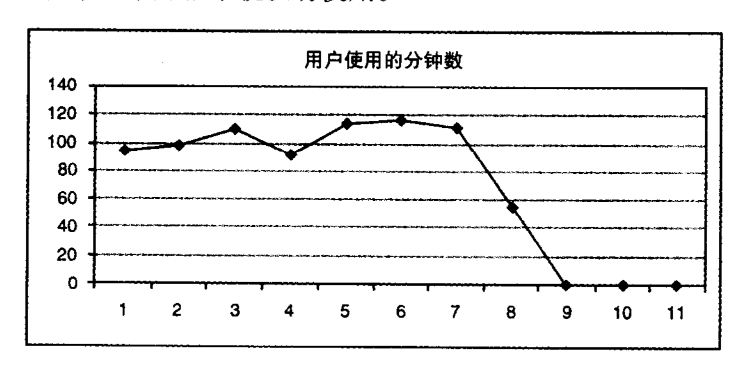


图 3-1　第 8 个月的使用下降预示着客户将在第 9 个月流失吗

这位用户看来适应这样的模式：在用户要放弃使用该项服务之前，有一个月的使用率会下降。但是表面现象具有欺骗性。注意该客户每天使用的分钟数而不是每月的使用情况，可

以发现该用户一直以一种固定的比率在使用这项服务直到那个月（第8个月）中旬的某一天，然后完全停止，大概因为在那一天，他（或她）开始使用一项竞争对手的服务。假定的使用率下降阶段实际上并不存在，因此当然不能提供保留客户的最佳时机。实际上最主要的线索似乎就是拖后的那一段。

图3-2显示了另一个由聚集引起混乱的例子。10月份的销售额似乎较8月和9月有所下降。图示数据来自仅在白天有销售活动的一个企业，这个时候金融部门也在营业。因为2003年10月的双休日和节假日都比较集中，所以10月份的交易日较8月和9月少一些，这就是10月份整体销售额下降的原因。

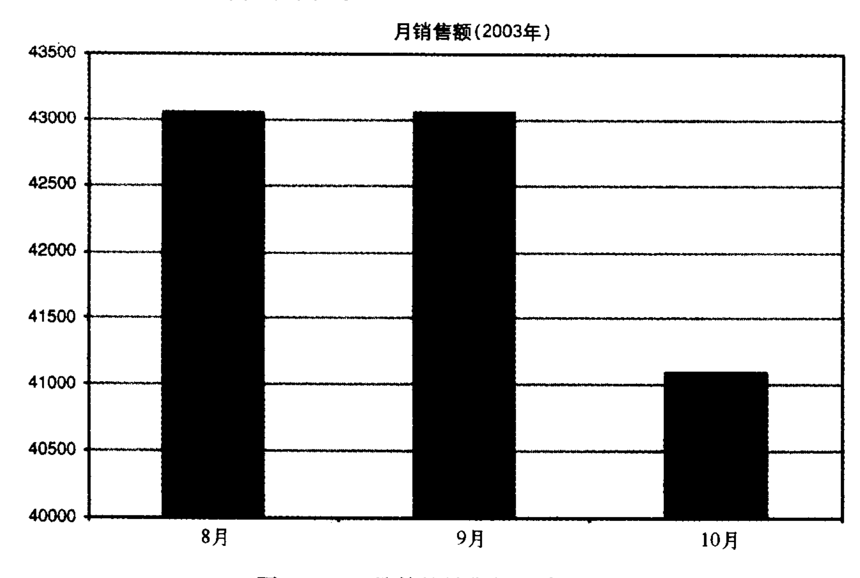


图3-2 10月份的销售额下降了吗

在前面的例子中，聚集（aggregation）运算引起了混乱，其实如果聚集层次不合适也会导致混乱。一个案例是，慈善机构提供的数据显示，捐赠人响应捐赠请求的可能性和捐赠数量成反比关系，即那些非常可能响应的人捐赠小额的支票。这种违反直觉的发现是慈善机构每年向支持者发出大量请求的结果。设想有两位捐赠人，每位计划向慈善机构捐赠500美元。一位响应1月份的请求，递送了一张500美元的捐献，并且把其他请求信扔到垃圾桶。另一位响应5次请求，每次捐赠100美元。在他们的税后年收入中，两位捐赠人都汇报捐赠了500美元，但是从个人行为层次来看，第二位看起来属于“更可能做出响应者”。当聚集运算以年为计算层次时，这种结果差异就消失了。

3.1.2 获取真实但无用的知识

获取无用知识的情况，虽然不像获取不真实的内容那样危险，但却是非常普遍遇到的现象。

1. 获取已知的知识

数据挖掘应该提供新的信息。数据中很多非常清楚的模式表示已知的知识。到退休年龄的人倾向于不响应那些退休储蓄计划。住在没有送货上门服务地区的人，不会订阅报纸，即使他们可能响应订阅服务，服务也不会开始。同样，住在没有移动信号发射塔地区的人也不

会买移动电话。

通常，很清楚的模式反映了商业规则。如果数据挖掘“发现”使用匿名呼叫业务的人也拥有呼叫号，这可能因为匿名呼叫业务仅仅是包含呼叫号的捆绑服务业务中的一部分。如果在一些特殊的地区没有这类产品销售，可能就不会在那里提供这种服务。我们曾经见到很多这样的发现，不是这些模式没有意义，只是它们的强度可能使某些不明显的模式黯淡。

获取已知的知识确实可以给我们一个有用的提示，从技术角度来说，这表明数据挖掘工作确有成效，而且数据本身也已经相当精确，这是非常令人鼓舞的。如果数据和所应用的数据挖掘技术足以发现已知正确的事实，据此可以相信其他发现也可能为真。数据挖掘也经常揭示应该知道但迄今还不知道的事情，例如：退休人员对退休储蓄存款账户的响应可能性不大。

2. 获取不能使用的知识

数据挖掘也时常揭示真实的和事先不知道的某些关系，但是仍然很难利用它们。有时候问题在于规章的限制：客户的无线呼叫模式也许间接表明某种有线长途通信业务包之间的密切关系，但是一个同时提供这两种业务的公司可能不被允许利用这个有利条件。类似地，客户的信用历史能够预言未来的保险索赔，但是规则可能禁止基于这一点做出保险决策。

另一些时候，数据挖掘发现重要的结果可能不在公司可控制范围之内。一种产品可能更适合某些气候，但是我们很难改变气候。可能由于地形原因，在某些地区的某种服务会很差，但这也很难改变。

提示：有时，缺乏想象会使新的信息看来是无用的。关于客户流失的案例研究极有可能表明客户要离开的最强信号是获得客户时的方式。回过头来修改现有客户的获取方式已经太迟了，但是这并不说明信息是无用的。通过改变不同获取渠道的组合来减少未来流失（future attrition），转向那些带来持久客户的渠道，可以减少未来客户的流失。

数据挖掘方法论的目的是避免获得不真实的知识，以及任何没有用的知识。更积极的理解是，数据挖掘的目标是确保数据挖掘得到稳定的模型，以便将该模型用于要解决的商业问题。

3.2 假设测试

假设测试（hypothesis testing）是整合数据到公司的决策制定过程（decision-making process）的最简单方法。假设测试的目标是证实或者反驳预想观点，是几乎所有数据挖掘工作的一部分。数据挖掘人员经常在各种方法之间来回反复，先是对观察到的行为（通常在商业专家的帮助下）给出可能的解释，并且抽取数据，分析假设的合理性，然后让数据给出要测试的新假设。

假设测试是科学家和统计学家惯于花费心血研究的事情。假设是一种解释，它的正确性可以由分析数据来检验。这些数据或者仅仅由观察收集，或者由实验生成，比如测试邮寄。当结果显示，用于指导公司市场行为的这些假设是不正确的时候，假设测试是最有价值的。例如，假定公司的广告是基于某个产品或服务的目标市场的许多假设以及响应本性，那么这些假设是否被实际的响应证实就非常值得测试。一种方法是对不同的广告使用不同的热线电话号码，记录每个响应者拨打的号码，然后将所收集的通话信息与广告最初期望影响的人群

进行比较。

提示：每次公司寻求客户的响应时，不管是通过广告或者其他的直接交流方式，都有机会收集信息。沟通方式的微小改变，如包括一种能够识别用户反馈渠道的方法，都可能大大增加所收集数据的价值。

假设测试本质上是不确定的，因此用“方法论”这个术语也许有点不恰当。然而，这个过程还是有一些可以确认的步骤，其中第一个也是最重要的步骤就是产生用来测试的好主意。

1. 产生假设

产生假设的关键在于从公司上下获得不同的输入数据，如果可能，最好从公司外部也获得一些数据。通常，开始这个认识过程的全部所需是清楚地表述问题本身，特别是以前没有认识到是一个问题的事情。

经常发生的现象是：用于评价公司业绩的度量没有捕捉到某个问题，所以这个问题一直没被注意到。如果公司总是基于每月的新销售量来测试销售能力，销售人员可能永远不会去考虑“新客户保持活跃的时间”或者“他们在公司与客户保持关系的问题上花费了多少”这类问题。然而当被问及这些问题时，销售人员可能意识到，某些客户行为是由于市场营销与客户之间的距离太远而错过了。

2. 测试假设

考虑下面的假设：

- 经常在外的人对移动电话每分钟的价格敏感度比其他人低。
- 有孩子在上中学的家庭更有可能响应家庭抵押贷款产品。
- 业务中心的挽留处正在挽留那些本可能回头的客户。

必须将这些假设以一种合适的方式在现实数据上测试。依据这些假设的不同，或许意味着要去解释简单查询返回的单个值，或在由购物篮分析（marketing basket analysis）产生的一堆关联规则中淘汰，或确定回归模型产生的关联的意义，或者设计对照实验（controlled experiment）等。在所有这些情况下，必须仔细认真地考虑，以保证结果在意外的情况下没有偏差。

正确评价数据挖掘的结果需要具有分析和商业两方面的知识。当不是由同一个人来处理这两方面的事情时，就需要进行交叉合作来充分利用新信息。

3.3 模型、建立简档和预测

假设测试当然有用，但有时还不够。本书下面介绍通过建立模型获取新知识的数据挖掘技术。

在通常状况下，模型是对某些事情的一种解释或者描述：它们能很好地反映现实世界，可用于对现实世界的推测。人类一直在有意或无意中利用模型。比如有两个餐馆，其中一家有白色的桌布，每个桌子上摆有鲜花，另一家用的是塑料贴面桌子，桌上摆着塑料花；你会推测前者比后者更贵，就是基于头脑中的模型进行推理的。当你走出餐馆步入店铺，关于这个镇的印象模型又一次留在了你的脑海中。

数据挖掘都是关于创造模型的问题。正如图3-3所示，通过使用一个输入数据集，模型会输出一个结果。用于创建模型的数据集称为模型集（model set）。当模型应用于新数据时，称为得分集（score set）。模型集由如下3部分组成，本章后面将给出详细讨论：

- 训练集用于建立模型集；

- 验证集[⊖]用于选出模型集中最好的一个模型；
- 测试集用于确定模型在未使用数据上的工作情况。

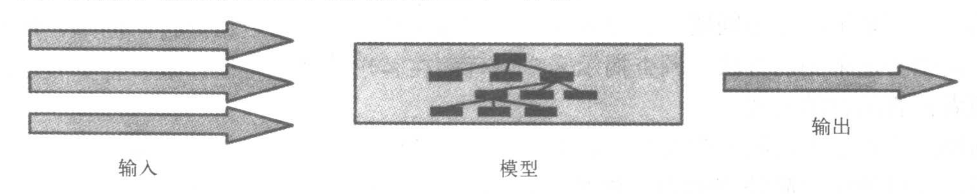


图 3-3 模型利用输入产生输出

数据挖掘技术可以为 3 类任务构造 3 种模型：建立描述性简档、建立定向简档（directed profiling）和预言（prediction），当然，它们之间的区别并不总是明显的。

描述性模型描述数据中存在什么，其输出通常是一个或多个表、数值或者图，解释当前正在发生的事情。假设测试经常会产生描述性模型。另一方面，建立定向简档和预测在模型建立初期都有一个预期的目标，它们之间的差别与时间帧有关，如图 3-4 所示。在简档模型中，输出和输入位于同样的时间帧，而在预言性模型中，目标位于下一个时间帧。预测意味着从一个时期的数据中发现模式，用来解释未来一段时间的结果。之所以强调简档和预测的区别，是因为这隐含了建模方法论，尤其是在创建模型集的时间处理上。

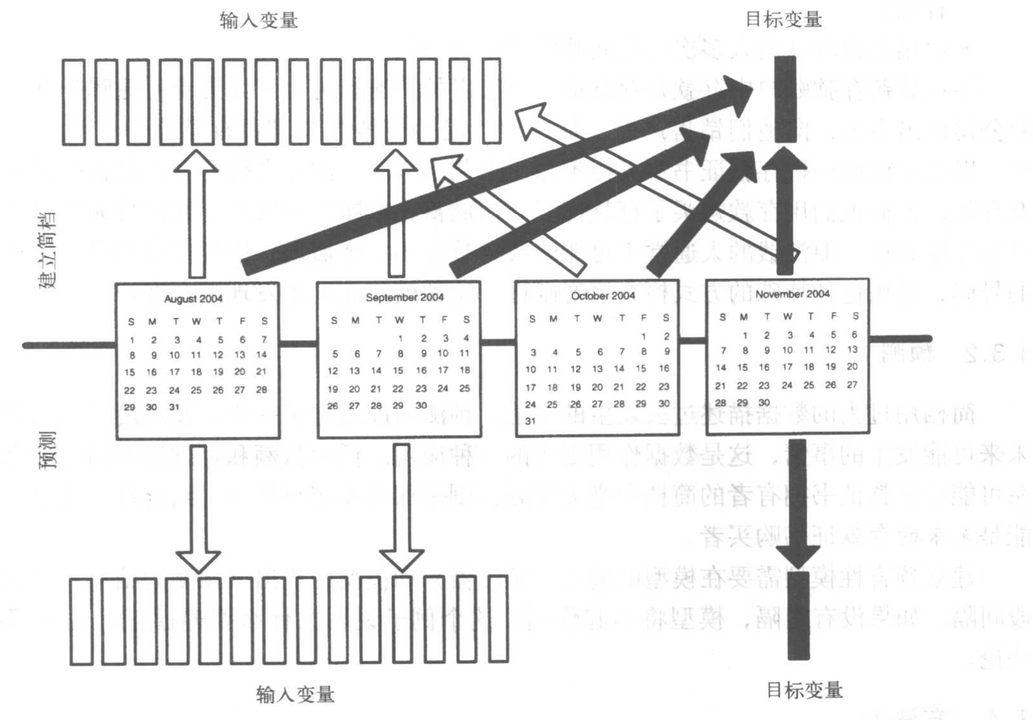


图 3-4 建立简档和预测的区别仅仅在于输入变量和目标变量的时间帧

⊖ 第 1 版中把模型集的三个部分称为训练集、测试集和评价集（evaluation set）。作者仍然喜爱这种术语，但是现在数据挖掘界的标准用法是训练集/验证集（validation set）/测试集。为避免混淆，本版采用了训练集/验证集/测试集的命名方法。

3.3.1 建立简档

建立简档是解决很多问题的常见方法，无需包含任何复杂的数据分析。例如，调查是建立客户简档普遍采用的方法。调查揭示客户或者潜在客户的外表特征，或者至少表明被调查的响应者回答问题的方式。

简档经常基于人口统计学变量，如地理位置、性别和年龄。因为广告也是基于同样的一些变量，人口统计学简档能够直接转换为媒体策略。简单的简档可用于设置保险费，一个 17 岁的男士比一个 60 岁的婆婆在汽车保险方面的费用更多。类似地，简单的人寿保险政策条款的申请表也询问客户的年龄、性别以及吸烟情况，除此之外，问得不多。

建立简档尽管被认为是一种强有力的工具，依然存在很大的局限性。一个缺点是不能区分因果关系。只要简档是基于熟知的人口统计学变量，这一点关系不大。如果男士购买的啤酒多于女士，我们没有必要惊奇喝啤酒是否是男性化的原因。看来，假设联系是从男士到啤酒似乎更可靠，反之则不行。

对于行为数据，因果关系的方向通常并不总是这样明显。考察下面来自实际数据挖掘项目的两个实例：

- 购买存款证书（certificates of deposit，CD）的人在储蓄存款账户中只有一点钱或者没有钱；
- 使用语音邮件的人多次短时间呼叫自己的号码。

不在储蓄存款账户中存款是存款证书拥有者的普遍行为，正如男士普遍喝啤酒一样。啤酒公司挑出男士，向他们销售产品，那么银行是否应该找出储蓄存款账户中没有存款的客户，以便向他们销售存款证书？大概不会。推测起来，存款证书拥有者在储蓄存款账户中没有存款，大概他们用存款购买了存款证书。而储蓄存款账户中没有存款的普遍原因可能是客户根本没有钱，但没钱的人通常不可能购买存款证书。类似地，语音邮件用户多次呼叫自己的号码，是用这种特殊的方式检查语音邮件。这种模式无助于发现潜在客户。

3.3.2 预测

简档用过去的数据描述过去发生的事情。预测向前迈进了一步。预测用过去的数据预测未来可能发生的事情，这是数据作用更大的一种应用。低存款额和存款证书拥有者之间的联系可能在存款证书拥有者的简档中毫无用处，倒是那些高额存款者（结合其他线索）极有可能是未来的存款证书购买者。

建立预言性模型需要在模型的输入（预测器）和模型的输出（预测的结果）时间上有一段间隔。如果没有间隔，模型将不起作用。这个例子说明为什么要遵循正确的数据挖掘方法论。

3.4 方法论

数据挖掘的方法论包括 11 步。

1）将商业问题转换为数据挖掘问题。

2）选取合适数据。

3）设法理解数据。

4）创建模型集。

5）修复数据问题。

6）变换数据，获取信息。

7）建立模型。

8）评估模型。

9）部署模型。

10）评估结果。

11）重新开始。

如图 3-5 所示，数据挖掘进程最好视为一组交叉的网状循环而不是一条直线。各步骤之间确实存在一个自然顺序，但是没有必要或苛求完全结束某个步骤后才进行下一步。后面几步中获取的信息可能要求重新考察前面的步骤。

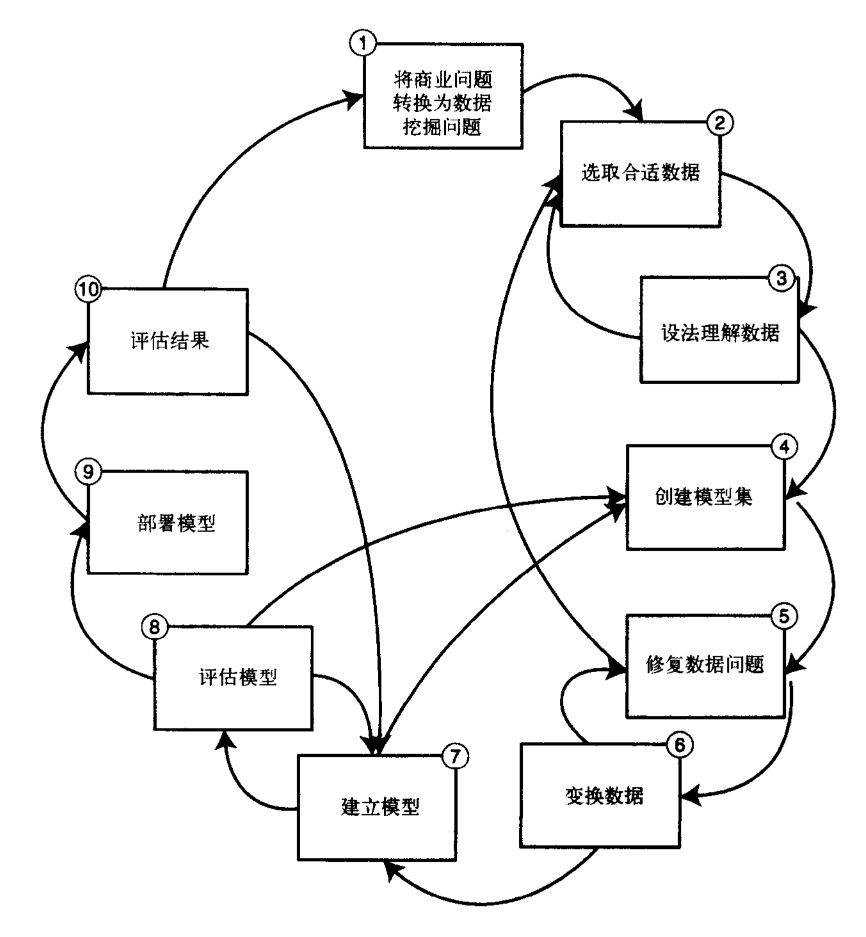


图 3-5 数据挖掘不是线性过程

3.4.1 第一步：将商业问题转换为数据挖掘问题

《爱丽丝漫游奇境》中精彩的一幕是爱丽丝向柴郡猫问路的那一段：

“请你告诉我，离开这里应该走哪条路?”

“这要看你想上哪儿去，”猫说。

“去哪里，我不大在乎。”爱丽丝说。

“那你走哪条路都没关系。”猫说。

“只要能走到一个地方。”爱丽丝又补充了一句。

“哦，那行，”猫说，“只要你走得很远的话。”

柴郡猫可能添加了假设：如果无法确定目的地，便永远无法分清你是否走了足够远。真正的数据挖掘项目的目标是为给定的商业问题提供解决方案。特定项目的数据挖掘目标不应该泛泛陈述，如：

- 更好地理解客户行为
- 发现数据中有意义的模式
- 获取有意义的信息

这些目标都很有价值，但即使达到了这个目标也很难测量。难于测量的项目也就难以评价其价值。无论什么情况下，泛泛的目标应该拆分成具体的目标，以便于监控实现进程。更好地理解客户行为可以拆分为以下具体的目标：

- 识别不可能再次订阅的用户；
- 设计话费套餐，减少居家商业客户（home-based business customer）的流失；
- 基于滑雪倾向给所有客户排序；
- 假如停止销售葡萄酒和啤酒，列出面临销售风险的产品。

这些具体的目标不仅易于监控，也易于转换为数据挖掘的问题。

1. 什么是数据挖掘问题

商业问题转换为数据挖掘问题时，应该表示为第1章介绍的6类数据挖掘任务之一的形式：

- 分类
- 估计
- 预测
- 关联分组
- 聚类
- 描述和建立简档

这是使用本书描述的技巧可以完成的一些任务，尽管没有哪个单独的数据挖掘工具或技巧对所有任务同样适用。

前三项任务：分类、估计和预测是定向数据挖掘（directed data mining）的例子。关联分组和聚类是非定向数据挖掘（indirected data mining）的例子。简档既可能是定向也可能是非定向的。定向数据挖掘总是有一个目标变量，表示分类、估计和预测的事情。建立一个分类器的过程可以从一系列预定义类的集合和已经正确分类的记录样本开始。类似地，建立估计器的过程以历史数据开始，其中目标变量的值是已知的。建模的任务就是发现用于解释目标变量已知值的规则。

在非定向数据挖掘中，没有目标变量。数据挖掘的任务是发现不依赖任何一个变量的总模式。非定向数据挖掘的最普遍形式是聚类，即不考虑哪个变量是最重要的情况下，发现相似记录的分组。非定向数据挖掘本质上是描述性的，因此经常用于建立简档，但是决策树等

定向技术对建立简档也很有用。在机器学习文献中，定向数据挖掘称为*有指导的学习*（supervised learning），非定向数据挖掘称为*无指导的学习*（unsupervised learning）。

2. 如何应用挖掘结果

这是确定如何把商业问题很好地转化为数据挖掘问题时最重要的问题。不可思议的是，最初经常的回答是："我们不能确定"。有一个答案是至关重要的，即不同的预期应用要求不同的解决方案。

例如，很多数据挖掘的约定是为改善客户的持久力而设计。这种研究的结果可以应用在以下任意一种情形：

- 抢先联络高风险/高价值的客户，并提供优惠待遇，使他们乐意留下；
- 改变获取渠道的组合，以期带来最忠诚的客户；
- 预报未来数月的客户群体状况；
- 改变产品，修正使客户流失的不足之处。

这些目标中的任何一个都与一个数据挖掘过程有关。通过电话销售或直接邮寄活动联络现有的客户，意味着除了确认风险客户之外，还能够了解他们为什么处于风险状态，因此可以构想有吸引力的优惠政策；并且了解他们什么时间将处于风险状态，因此可以在非常恰当的时间联络他们。预报意味着除了确认哪些现有客户可能离开之外，还可以确定可能增加多少新客户，以及他们可能停留多长时间。预测新客户可能停留多长时间的问题，通常包含在商业目标和预算中，不是预言性建模的问题。

3. 以何种方式交付结果

数据挖掘计划可能有几种不同类型的交付方式。如果最初的数据挖掘目标是获得市场了解，交付方式通常是一份充满图或表格的报告或介绍。如果该项目是一个一次性的概念验证或者小规模试验计划，交付方式可能是包含在将来的销售试验中获得不同待遇的客户列表。当数据挖掘项目是正在进行的客户关系管理分析的一部分时，交付方式极可能是一个或者一系列计算机程序，可以定期运行，给客户群体中预定义的子集打分，并且随时间和另外的软件一起管理模型和评分。交付方式可能影响挖掘的结果。假如目标是使销售管理人员产生深刻印象，只给出销售测试产生的客户列表是远远不够的。

4. 商业用户和信息技术的角色

正如第2章所述，获得以上问题的正确解决方案的惟一方法是让商业问题的所有者参与其中，判断将如何应用数据挖掘的结果，让IT员工和数据库管理员参与，来判断如何交付结果。通常同时在企业内部和合适的外部领域广泛收集数据是非常有用的。我们建议把企业各部门的代表集中到一起，而不是分别单独会见他们。用这种方式，具有不同知识领域的人和专家有机会相互交流各自的思想。所有这些磋商的目标是获得所讨论的商业问题的清晰陈述。最终的商业问题陈述必须尽可能具体。"确认10 000位重量级客户极有可能在未来的60天之内流失"比"为所有客户的流失可能性打分"更好。

误解商业问题：一个有警诫意义的事件

作为数据挖掘者，受一个大的包装消费品厂商的委托，作者曾经担任顾问，参与分析超级市场忠诚卡的数据。理顺这些关系，可以了解一点关于超级市场行业的信息。一般来说，超级市场不关心客户购买可口可乐还是百事可乐（除非其中一个品牌正在促销，因而带来更高的利润），只要客户购买软饮料就行。供应商则非常关心销售了哪种品牌，争取管理商店

整体分类的机会。而销售分类的管理者，能够控制自己或者竞争对手的产品销售。客户希望展示使用忠诚卡改善分类管理的能力。挑选出来用于展示的分类是酸奶酪，因为根据超级市场的标准，酸奶酪是高利润产品。

正如我们理解的一样，商业问题的目标是确定喜欢酸奶酪的人。为创建一个目标变量，我们根据一年内酸奶酪的全部购买量，把忠诚卡客户分成高、中、低酸奶酪关联分组，并且根据他们购买酸奶酪的花费在全部消费额中的比例把客户分为高、中、低客户。把在两种度量中都属于高级客户的消费者标记为喜欢酸奶酪的人。

交易数据需要经历很多变换，最终才能变为客户特征。输入变量包含以下项：在一天的不同时间和不同分类中购买酸奶酪的次数以及消费金额占任何购买分类、购买频率、平均订货规模和其他行为变量的比例。

使用这些数据，我们建立了一个模型，给每一位客户一个酸奶酪喜爱程度得分。拥有这样的得分后，当可能的酸奶酪喜爱者付账离开时，就可以打印关于酸奶酪的优惠券，尽管他们这次可能没有购买酸奶酪。该模型甚至可以确认好的潜在客户，尽管这些潜在客户还没有与内部的酸奶酪喜爱者有任何联系，但是如果给予优惠券，他们有可能立刻成为酸奶酪喜爱者。

该模型很令人鼓舞，我们对此非常满意。然而委托人却失望了。“但是，谁才是喜爱酸奶酪的人?”委托人问道，“在该模型中得高分的人”未必就是委托人想要的回答。客户要寻求的是类似“喜爱酸奶酪的人是年龄在 x 和 y 之间的女士，她们住在平均家庭收入介于 M 和 N 之间的地区”。像这样的描述可以用于决定在哪里投放广告，以及如何制作有创意的广告内容。由于我们的模型建立在购物行为而不是人口统计学基础上，因此不能满足客户的要求。

在这些讨论中，数据挖掘者的角色主要是确保商业问题的最终陈述是那些可以顺利转换为数据挖掘问题的陈述。否则，世界上最好的数据挖掘工作可能被用于解决一个错误的商业问题。

通常把数据挖掘表现为一个技术问题，即找到一种模型来解释目标变量到输入变量群体之间的关系。这类技术任务对大多数数据挖掘工作来说确实非常重要。但是在精确确定目标变量和确认合适的输入变量之前，不能尝试这件事情。这反过来依赖于对所讨论的商业问题的良好理解。就像前面讲的故事“误解商业问题”一样，若不能正确地把商业问题转换为数据挖掘问题，会导致我们试图避免的危险发生，即获取的内容是真实的，但却没有用处。

要完整地把一个商业问题转换为数据挖掘问题，建议参考我们的同事 Dorian Pyle 所著的 *Business Modeling and Data Mining* 一书，书中对“如何发现数据挖掘受益最大的商业问题”以及“如何为数据挖掘明确地表达出这些问题”方面给出了详细的建议。作者在此仅提醒读者，在进行实际的数据挖掘过程之前，考虑两个重要的问题：结果将如何应用？结果将以何种形式交付？从第 1 个问题的答案到第 2 个问题的答案，仍然需要一个很长的过程。

3.4.2 第二步：选取合适数据

数据挖掘需要数据。在所有可能的情况中，最好是所需数据已经存储在共同的数据仓库中，经过清理，数据可用，历史精确并且经常更新。事实上，它们经常以不兼容的形式散列在各种操作系统平台的计算机上，这些计算机之间运行着不同的操作系统，通过不兼容的桌

面工具来访问。

当然，根据问题和产业不同，有意义的和可用的数据源（data source）也不同。一些有益数据的例子如下：

- 保质期内索赔数据（包括固定格式的数据和自由文本字段）
- 销售点数据（包括环形码、提供的优惠券、折扣）
- 信用卡收费记录
- 医疗保险索赔数据（medical insurance claims data）
- 网络日志数据
- 电子商务服务器应用程序日志
- 直接邮寄响应记录
- 呼叫中心记录，包括呼叫中心人员撰写的备忘录
- 打印机的运行记录
- 机动车登记记录（motor vehicle registration record）
- 安放在机场附近的社区中的扩音器产生的噪声分贝
- 电话呼叫详细记录
- 调查响应数据（survey response data）
- 人口统计和生活方式数据
- 产、供、销数据
- 每小时天气情况（风向、风力、降雨量）
- 人口普查数据

一旦商业问题完成公式化，就可以构造一个拥有最佳数据的数据列表。对于研究现有客户来说，需要包括从他们成为客户的那一刻起的数据（数据获取渠道、获取日期、最初的产品组合、最初的信用评分，等等），描述他们当前状态的类似数据，以及客户保有期内积聚的行为数据。当然，不可能从数据列表中找到所有数据，但是最好从你想找到什么数据着手。

有时候，开始一项数据挖掘项目之初并没有一个特定的商业问题。公司意识到从收集的数据不能得到很高的价值，开始思考通过数据挖掘是否会使数据更有价值。这类项目成功的秘诀在于把它转换为一项为解决特定问题而设计的项目。第一步是探索可用的数据，写出候选商业问题列表，邀请商业用户创造一个非常长的待选列表，然后将其简化成少数可以达到的目标，即数据挖掘的问题。

1. 什么数据可用

寻找数据的首选是公司的数据仓库。仓库中的数据是已经清理、校验过，并且已把多种数据源整合。单个数据模型有希望保证类似命名的字段在整个数据库中有同样的意义和相容的数据类型。公司数据仓库是历史仓库，可以增加新的数据，但历史数据永远不再改变。因为它为决策支持而设计，数据仓库提供详细的数据，可以聚类到正确的层次，以利于数据挖掘。第 15 章更加详尽地讲述数据挖掘和数据仓库之间的关系。

惟一的问题是，在许多机构这样的数据仓库并不存在，或者存在一个或者多个数据仓库，但是不能达到上述要求。在这种情况下，数据挖掘人员必须从各部门数据库和操作系统内部寻找数据。操作系统是为完成特定的任务而设计，如索赔处理、呼叫转换、订货登记或

者付账等，最初的设计目的是快速、准确地处理交易。不管数据的格式如何，目的是很好地适合特定目标，如果有历史数据的话，也可能存储在磁带存储器。也许需要大量的规章调整和编程工作，才能得到对知识发现有益的数据格式。

有时候，为了支持数据，可能要改变操作过程。我们知道有一位目录销售商希望分析客户的购买习惯，以便对新客户和持久客户进行不同的销售。不幸的是，在过去的6个月之内没有订购任何东西的人被从记录中例行清除。忠诚地使用目录在圣诞节购物，而在一年的其他时间没有购买东西的大多数人都没有被识别，事实上是*不可识别*，一直到公司开始保存历史数据，这种情况才被改变。

在很多公司中，想确定什么数据可用是非常困难的，因为档案资料经常丢失或过期。通常而言，没有任何人能够提供所有问题的答案。想确定什么数据可用，需要仔细查阅数据，与用户和数据库管理员交流，或者仔细检查已有的报告。

警告： *使用数据库文档和数据字典作为指南，但是不要视之为一成不变的事实。在表中定义的字段或者在文档中提到的字段，不能说明字段的存在，这实际上对所有客户都可用。*

2. 多少数据够用

遗憾的是，这个问题没有一个简单的答案。答案依赖于所使用的特别算法、数据复杂程度、可能结果的相对频繁程度。统计学家已经花费数年时间开发测试手段，以确定产生模型的最小模型集。机器学习研究人员花费很多时间和精力设计方法，使得训练集的一部分可以重用于验证和测试。所有这些工作忽略了重要的一点：在商业界，统计学家很缺乏，而数据很多。

任何情况下，如果数据缺乏的话，数据挖掘不仅有效性差，而且不大可能有用。当小型数据库中相当大量的数据掩盖了可探查到的模式时，数据挖掘最有用。因此我们建议使用足够多的数据，使得不会出现“足够大的样本集的规模是多少”这个问题。即使不采用数以百万计的预分类记录，我们一般也会采用数万条记录，以保证训练集、验证集、测试集都包含上千条记录。

在数据挖掘中，数据越多越好，但是有几个忠告。第一个忠告是关于模型集的规模和密度（density）的关系。密度指的是利益输出的普遍性。通常，目标变量代表相对稀有的事情。潜在客户很少响应直接邮寄广告，信用卡持卡人很少欺诈，报纸订阅者很少在指定的某个月取消订阅。正如本章后面（创建模型集部分）要讨论的，在创建模型的过程中，最好使每个模型集的各个输出数量大致相同。一个较小的均衡样本比含有稀有输出比例极低的较大样本更可取。

第二个忠告与数据挖掘人员的时间有关。当模型集大到足以建立一个很好的、稳定的模型，再增大模型集反而会有负面的影响，因为在更大的数据集上，每件事情都要花费更长的时间。由于数据挖掘是一个反复的迭代过程，如果建模过程的每一步运行时间不是数分钟而是数小时的话，等待时间可能变得非常长。

测试用于建模的样本规模是否足够大的一个简单办法是，试着加倍样本数，测试模型精度的变化。如果用大的数据样本创建的模型比使用小的样本创建的模型要好得多，那么小的样本不够大。如果没有任何变化，或者仅有微弱的变化，那么原来的样本可能是合适的。

3. 需要多少历史数据

数据挖掘使用过去的数据预测未来。但是需要多久以前的数据才合适呢？这是另一个没

有简单答案的简单问题。要考虑的第 1 个问题是季节性，大多数商业活动都表现出一定程度的季节性：第 4 季度的销售上升，休闲旅游在夏季升温，等等，应该有足够的历史数据来捕捉这类周期性事件。

另一个方面，由于销售条件的改变，太久远的数据可能无益于数据挖掘，特别是受一些外部事件的影响，如政治制度的调整变化等，就更会出现这种情况。对很多面向客户的数据挖掘来说，2～3 年的历史数据就是合适的。即使在这种情况下，关于客户关系建立初始时的数据常常是很有价值的，例如，初始渠道是什么，初始的优惠条件是什么，客户最初如何支付，等等。

4. 需要多少变量

没有经验的数据挖掘者经常匆匆忙忙删掉一些看来不太有意义的变量，保留仔细挑选后他们认为很重要的少数几个变量。数据挖掘的方法提倡让数据本身揭示哪些变量重要，哪些不重要。

最初忽略的变量，如果结合其他变量使用，经常被证明有预言性价值。例如，在一个信用卡发行者的客户利润模型中，从来没有包含过现金预付，通过数据挖掘发现，只在 11 月和 12 月使用现金预付的客户是非常有利可图的。这些人大概非常谨慎，他们大部分的时间避免在高贷款利率情况下借款（谨慎使他们更不大可能比习惯性使用现金预付的客户拖欠还贷），但是他们过节期间需要额外的一些现金，并且愿意付出昂贵的利息来得到它。

最后的模型通常基于几个变量，但是这些变量经常是结合一些其他变量而产生的，往往开始时并不明显，最后却发现很重要。

5. 数据必须包含什么

数据至少要包含所有可能的有意义的结果的例子。在定向数据挖掘中，目标是预测特定目标变量的值，有一个由预分类数据构成的模型集至关重要。为了区别可能拖欠贷款的人和不可能拖欠贷款的人，需要从每个类取出上千个例子建立模型，以便于区分。假如有新的申请者，把他或她的申请和过去的客户比较，可以采用基于存储信息的推理进行直接比较，也可以采用源于历史数据的规则或神经网络进行间接比较。如果发现新的申请人“看来”会像过去有过拖欠的人一样，他就会被拒绝。

隐含在这一描述中的观点是，我们可能知道过去发生的事情。从过去的错误中吸取教训，首先应该认识到所犯的错误，这种事情并不总是可能的。一个公司不得不放弃使用定向知识发现来创建保险索赔欺骗模型的尝试，尽管他们怀疑其中的一些索赔具有欺骗性，但是无法说明到底哪些是欺骗性的。没有一个包含明确标识为欺骗性还是合法性的保险索赔训练集，就无法使用这些技术。另一个公司希望创建一个直接邮寄响应模型，但是仅支持对过去活动响应者的数据，没有包含非响应者的任何信息，所以也无法进行比较。

3.4.3 第三步：设法理解数据

在开始建立模型之前，无论花多少时间研究数据都是值得的。正是由于这个过程的重要性，在第 17 章中将会详细讨论这个主题。好的数据挖掘人员在很大程度上依赖于直觉（intuition）——比如在某种情况下能够猜出一个好的衍生变量可能是什么样的。培养这种直觉的惟一方法就是全身心地投入到陌生的数据集合中。这样你就有可能发现许多数据质量问题，受到更多启发，去问许多其他方式下不会想到的问题。

1. 检查分布状况

一个良好的开端是检查数据集中的每一个变量的直方图，并考虑这些直方图说明了什么问题。所有看上去不同一般的东西都要记下来。如果有一个州代码变量，代表加州的直方块是不是最高的？如果不是，为什么？是不是漏掉了某些州？如果是，这个公司在那些州没有商务活动合理吗？如果有一个性别变量，男性和女性的数量是不是差不多？如果不是，是否出乎意料？要注意检查每一个变量的范围。计数变量是否取成负值？变量的最高值和最低值看起来是合理的赋值吗？平均值是不是与中间值的差别很大？丢失了多少数值？变量计数与时间一致吗？

提示：一旦从新的数据源获得了数据文件，为了了解下一步要做的事情，为数据建立简档是一个好办法，包括获取每个字段的计数和概要统计、计算分类变量的不同取值的数量，如果需要的话，使用交叉表来显示数据，比如根据产品和地区来显示销售量。除了提供对数据的深入了解外，建立简档的训练有可能产生关于不一致性或者清晰度问题的警告标志，这些不一致性或者清晰度问题会破坏以后分析的有效率。

在研究数据库的初始阶段，数据可视化工具是非常有用的。图3-6显示的是2000年纽约州人口普查的一些数据。（这个数据集可以从www.data-miners.com/companion下载，网站上还有一些使用这个数据集的练习题。）黑色的条表示在那些县中使用木材取暖的家庭超过15%的村镇比例。（在纽约，村镇是县的下属机构，一个村镇可能包含一体化的乡村或者城市，也可能不包含。如Cortland镇属于Westchester县，并且包含了Croton-on－Hudson村，而Cortland市属于Cortland县，位于州的另一部分。）这个图是用Quadstone公司的软件生成的，从图中很容易看出，在纽约附近的城市化县中用木材取暖的家庭并不多，但是北部的乡下地区却非常普遍。

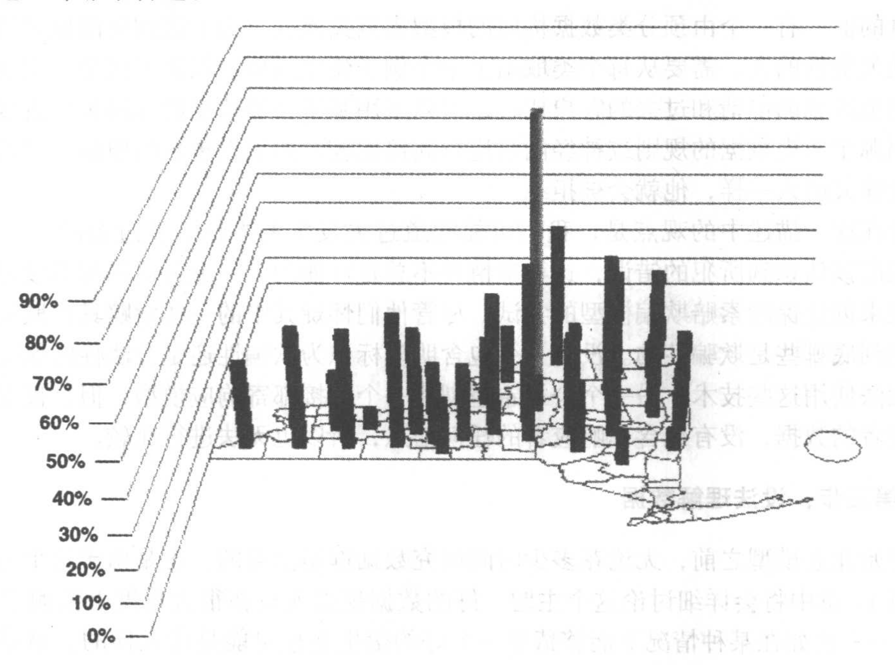


图3-6 纽约州以木材为主要加热原料在各县的盛行情况

2. 比较变量值及其描述

考察每个变量的取值，并将其与在可用文档中提供的关于该变量的描述进行比较，这样的比较经常可以发现描述是不准确的和不完整的。在一个杂货购买数据集中，被标记为项目计数的变量取了很多非整数值。进一步的研究发现，本应按照销售物品进行计件的一个条目，却是按照销售物品的重量来记录的。在由零售公司提供的另一个数据库中，有一个字段被描述为几个季度的总开销，不可思议的是，这个字段被用于预测目标变量，而不管客户是否从特定目录邮件下了订单。凡是没有下订单的客户在这个字段的取值为0，凡是下了订单的则会有一个大于0的取值。我们推测，这个字段事实上可能包含了从所关注邮件中得到的客户订单的值。无论如何，其中肯定不会包含已经记录的值。

3. 验证假设

利用简单的交叉表和可视化工具如散点图、条形图和示意图，可以验证关于数据的假设。与其他不同变量关联起来考察目标变量，可以理解诸如不同响应的渠道、不同市场的流失率或者收入按性别的差异等细节。如果可能的话，尽量通过直接从基层数据重新生成的方法来比较报告的概要数字是否准确。例如，如果报表生成的市场流失率为2%，那么可以通过计算一个月内取消业务的客户数量总和，看是否占总数的2%左右，以此来验证报告的正确性。

提示：从详尽的数据中设法重新计算已聚集的数值，是一项有启发性的训练。在试图解释这种差异的过程中，极有可能获取报告数据背后的、关于操作过程和商业规则的知识。

4. 提问问题

记下所有的与已有知识或者期望值不符的数据。对数据进行探究的一个重要内容就是要给数据提供者提一系列的问题。由于很少有用户会像数据挖掘者那样仔细对待数据，所以这些问题需要进一步研究。以下是对数据的初步探索中可能经常出现的问题：

- 为什么在新泽西州或马萨诸塞州没有汽车保险销售业务？
- 为什么一些客户在2月份活跃31天，而没有人在1月份活跃28天以上呢？
- 为什么有那么多客户出生在1911年？他们的年龄真的那么大吗？
- 为什么没有重复购买的实例？
- 合同开始日期晚于结束日期意味着什么？
- 为什么销售价格字段会出现负值？
- 活跃的客户怎么可能在“取消理由”字段有非空值？

这些都是我们在查询实际数据时遇到的真实问题。这些问题的答案有时会提供一些我们以前不知道的客户领域的知识。新泽西州和马萨诸塞州禁止汽车保险公司在设定比率时有太大的弹性，所以主要依靠价格竞争的公司就不愿意进入那些市场。有时，我们可以从这些问题中获知关于操作系统的一些特性，比如在登录数据时，如果对客户一无所知而系统却要求输入出生日期，很多人就会按下键盘上的“1”键，直到填满这个字段为止（如果一直按其他键，则输入的日期无效），此时输入的出生日期就是11/11/11，这就导致很多人的出生日期都是1911年11月11日。有时还会发现数据的一些严重错误，比如把2月份的数据误认作1月份的数据等。在最后一个实例中，我们还发现在提取数据的过程中存在漏洞。

3.4.4 第四步：创建模型集

模型集包含了在建模过程中用到的所有数据。模型集中有些数据用于发现模式，有些数

据用于验证模型是否稳定，有些用于评价模型的性能。创建模型集需要从多个数据源收集数据以构成客户特征标识，并为分析准备数据。

1. 收集客户特征标识

模型集就是一个表格或者多个表格的集合，其中每一行对应一个待研究的条目，字段则是与对建模有用的条目相关的所有内容。在用这些数据描述客户的时候，模型集中的行通常称为*客户特征标识*（customer signature）。从关系数据库中收集的客户特征标识往往需要连接多个数据表进行复杂的查询，然后通过其他的数据源进行扩充。

数据收集过程的一部分工作是从正确的汇总（summarization）层次上得到所有的数据，这样每一位客户对应一个值，而不是每次交易或者每个邮政编码对应一个值。这些问题将在第 17 章中继续讨论。

2. 创建平衡样本

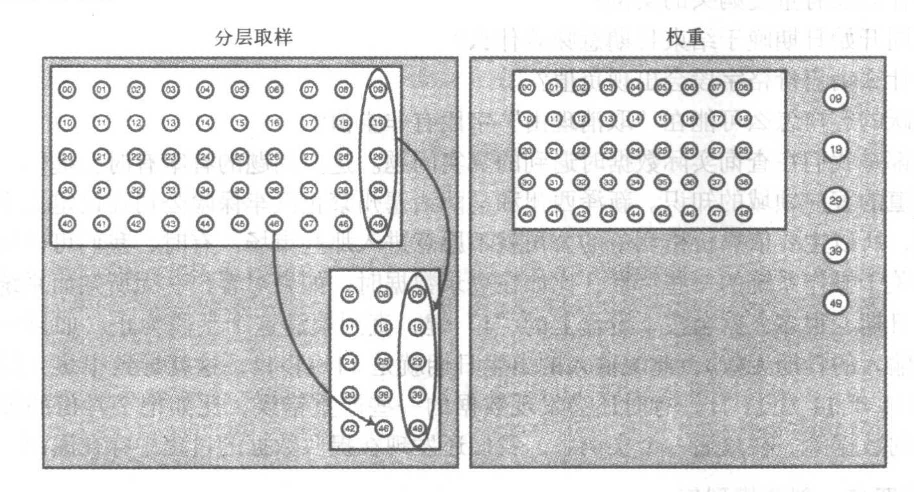
数据挖掘的任务经常会涉及学会区分不同的群体，比如响应者和非响应者，好的和差的，或者不同客户群的成员等。正如在下面的“不能轻易抛弃离群值”部分介绍的那样：在这些群体的成员数量大致相同时，数据挖掘算法的效果最佳。这种情况是不大可能自然出现的，事实上，成员不足的那些群体往往更值得关注。

在开始建模之前应该通过下面两种方法来平衡数据集：1）不同群体以不同的比例取样；2）添加权重因子，使得最大的群体与最小的群体成员的权重不同。

不能轻易抛弃离群值

在标准的统计分析中，抛弃那些远离正常范围的**离群值**是通用的惯例。但是在数据挖掘中，这些离群值或许正是我们寻求的，它们或许是在业务流程中出现的某种失误造成的假象，也许是某种难以置信的利润丰厚的市场机会。在这种情况下，我们不能轻易地抛弃，而是要逐步认识并理解这些离群值。

问题在于知识发现算法依赖于样本学习。如果没有足够的特定类或者行为模式的样本，数据挖掘工具将难以提供能够预测的模型。这种情况下，或许能够通过人工添加特例训练数据的方法来解决。



在结果稀有的情况下，有两种创建平衡样本（balanced sample）的方法

例如，一家银行想对可能参加私人银行计划的客户建立一个模型。这样的计划只是为了吸引那些最最富有的客户，即使在一个相当大的银行客户样本中，这样的客户也是罕见的。为了建立一个能够刻画这些富有客户的模型，应该创建一个训练集，使其包含50%的私人银行客户的交易历史，尽管他们在所有客户中占的比率不会超过1%。

另一个可能的办法是，选择给每一个私人银行客户赋予权重为1的值，而其他客户的权重值为0.01，这样高级客户的总权重才会和其他客户的权重基本相等（我们一般给的最大权重就是1）。

3. 包含多种时间帧

方法论的基本目标就是创建稳定的模型。这起码意味着，模型应该能够在任何时候和未来都运转良好。当模型集中的数据不都来自同一年的某个时间时，更容易出现这种情况。即使一个模型只是基于3个月的历史数据，模型集中不同的行也可能使用不同的3个月时段。解决问题的思路应该是从过去的数据进行概括产生模型，而不仅仅记录过去某一特定时间发生的事情。

基于单一时间段建立模型，就会增加以偏概全的风险。作者曾遇到一个有趣的例子，有人仅用某超市一周的销售数据建立了一个关联规则模型。关联规则的目的是，在给定了购物篮中的某些商品后，预测购物篮的另外一些商品。在这个例子中，所有的关联规则预测的结果都是鸡蛋。这种奇怪的结果是很少见的，因为后来我们发现这个模型集是基于复活节前一周的数据建立的。

4. 创建预言性模型集

当模型集用于预测时，还要考虑关于时间概念的另一个问题。尽管模型集应该包含多种时间帧，任何一个客户特征标识都可能包含预言性变量和目标变量之间的一个时间差异。时间总是被分为3个阶段：过去、现在和将来。在进行预测时，模型是使用过去的数据预测未来情况。

正如图3-7所示，这3个阶段都应该在模型集中表示。当然所有的数据都来自过去，模型集中的时间区间应该是“遥远的过去”、“不太遥远的过去”和“最近的过去”。预言性模型就是发现“遥远的过去”的某种模式，来解释“最近的过去”发生的结果。当模型被部署以后，就可以使用“最近的过去”的一些数据预测未来状况。

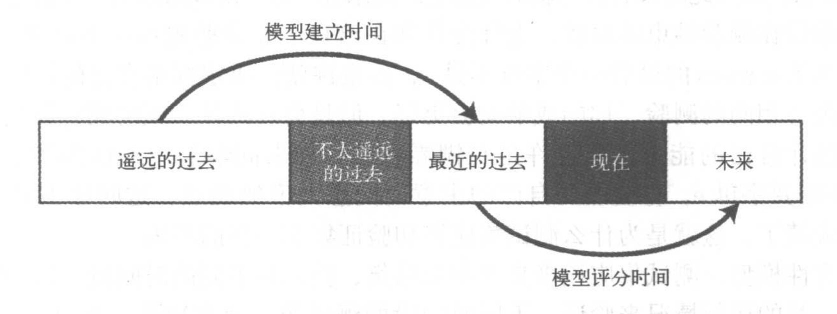


图3-7 过去的数据模拟来自过去、现在和未来的数据

至于为什么有些比较新即“不太遥远的过去”的数据没有在某个客户特征标识中得到使用，原因往往不太明显。这是因为对于目前应用的模型，最新的数据还没有作为输入应用到模型中，图3-8清楚地表明了这一点。

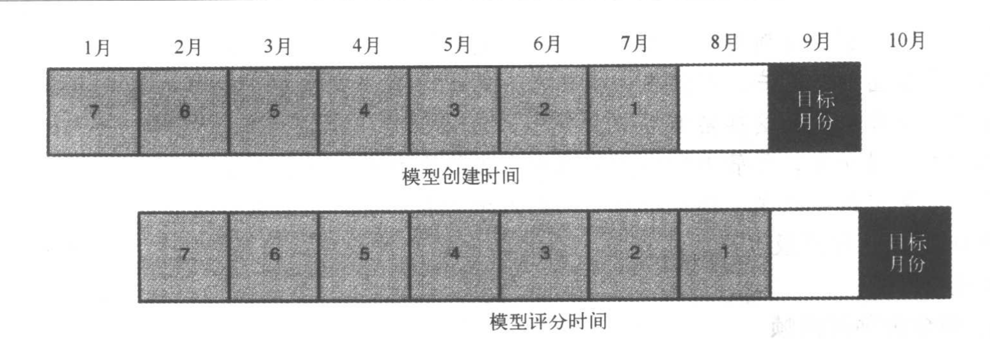


图 3-8　建立模型时间和使用模型时间的比较

假如有一个模型，使用6月份的数据（不太遥远的过去）来预测7月份（最近的过去）的情况，那么只有当8月份的数据可用的时候，这个模型才能够用于预测9月份的情况。可是，8月份的数据什么时候可用呢？事实上，在8月份可用是不可能的，因为8月份的数据正在生成。数据总会有，但可能在9月份的第一周还不行，因为数据需要收集、清理、装载、测试等。在许多公司中，8月份的数据到9月中旬甚至10月份才可以使用，但此时已经不再有人关心对9月份的预测了。对这个问题的解决方法，就是要在模型集中包含一个月的延迟时间。

5. **划分模型集**

在从合适的时间段获得了预分类的数据以后，方法论要求将其分为三个部分。第一部分是训练集，用于创建初始的模型。第二部分是验证集，用于调整初始的模型，使它更加通用，而不至于过多地依赖于训练集。第三部分是测试集，用于测试把模型用于未经训练的数据时可能的有效性。分成这样的三个部分是必要的，因为一旦某些数据在上述过程的某一步中使用过，其信息也就变成了模型的一部分，这些数据也就无法应用到下一步中，因而也就不能用于修正或者评价模型。

人们可能难以理解为什么测试集和验证集一旦用过之后就会“变质”。打个比方，如果你在上五年级，课堂上正在进行一个拼写测验。假如在测验要结束的时候，老师让你标出试卷上拼错的单词并给自己打一个成绩，你肯定给自己一个高分，但是你的拼写能力却没有得到提高。如果在一开始，你认为 tomato 的最后一个字母应该是 e，当你给自己评分的时候，就没有任何东西可以改变你的主意，因为此时没有任何新的信息进入到系统中来，所以必须有一个验证集。

现在，假设在测验结束的时候，老师允许你在给自己评分前先看一下周围同学的试卷。如果他们都认为 tomato 的最后一个字母不是 e，你也许就会决定标出自己的错误了。如果老师在第二天进行相同的测验，你的成绩就会更好。但是会好到什么程度呢？如果使用近邻同学的试卷来估计自己的能力，那也许是在糊弄自己。如果同学们一致认为和 tomato 一样，potatoes 也不需要字母 e，你会改变自己的主意而同意大家的观点，这时你就会高估自己第二次测验的成绩了。这就是为什么测试集应该和验证集不一样的原因。

对于预言性模型，测试集应该来自于和训练集、验证集不同的时间段。模型的稳定性要通过其月复一月的运行情况来验证。不同时间段的测试集，通常称作过期（out of time）测试集，是测试模型稳定性的一个好方法，尽管这样的测试集并不总是可用的。

3.4.5　第五步：修复数据问题

所有的数据都是“脏”的，所有的数据都会有问题。数据是否有问题要随数据挖掘技术

的不同而定。对某些技术而言，比如决策树、缺失值和离群值不会引起太多问题；对于另外一些技术，比如神经网络，就会有各种各样的麻烦了。正因为如此，一部分数据修复问题可以在相关技术的各章节中进行讨论，剩下的那些问题可以在第 17 章的“数据的黑暗面”一节中找到相应的内容。

接下来的几小节要讨论的是在数据修复中的一些共性问题。

1. 拥有太多数值的分类变量

诸如邮政编码、区县、电话听筒样式和职业编码等变量都是传递有用信息的，但那不是大多数的数据挖掘算法能够处理的方式。主要问题在于，尽管一个人住在哪里和做什么工作是很重要的预测因素，但传递这些信息的变量如此之多，而对于大多数的取值，数据中的样本又是如此之少，所以像邮政编码、职业这样的变量，连同它们所表达的有价值信息一起被抛弃掉了。

像这样的变量有两种处理方法，一是组合，将许多具有与目标变量近乎相同的关系的变量可以组合在一起；二是将它们替换为与邮政编码、电话听筒样式或者职业相关的有意义属性。可以将邮政编码替换为以下属性：邮政编码代表的中间本土价格、人口密度、历史响应率，或其他具有预言性意义的属性。将职业替换为相应职业的平均工资，等等。

2. 具有倾斜分布和离群值的数值变量

对于任何使用算术运算（比如，数值与权重的乘积和数值的和）的数据挖掘技术，倾斜分布（skewed distribution）和离群值（outlier）都会引发问题。在许多情况下，抛弃离群值记录是有意义的。而在另外一些情况下，最好把数值分成大小相等的范围，比如十分位数。有时候，转换这类变量的最佳方法是通过对取值求对数等方法将其取值范围进行缩减。

3. 缺失值

有些数据挖掘算法能够将“缺失”看做一个值并将其融入规则中。不幸的是，另外的一些算法不能够处理缺失值（missing value）。没有任何简单直接的解决方案能够保留变量原有的真实分布。抛弃所有带有缺失值的记录会引起偏差，因为这些记录并不是随机分布的。将这些缺失值替换为某些可能的值，比如平均值或者最常见的值，会增加一些虚假的信息。而将这些值替换为某些根本就不可能的取值就更糟糕了，因为数据挖掘算法不会识别年龄变量的取值为 -999 这样的情况，算法会继续进行并使用该值。

当缺失值必须被替换时，最好的方法是通过创建模型转移它们，这个模型把该缺失值作为目标变量。

4. 含义随时间变化的值

当数据来源于过去的不同时间点时，同一字段的同一取值所表示的含义随时间而变化的情况是很常见的。信用等级为“A”总是最好的，但是对应于层次 A 的具体的信用得分范围是随时间不断变化的。要恰当地处理这样的问题，就需要一个设计良好的数据仓库，在数据仓库中记录这些含义变化，并定义一个含义不随时间变化的新变量。

5. 不一致数据编码

当针对同一主题的信息来源于多个数据源时，相同的数据可能会有不同的表示方法。如果不能很好地把握这些不同的表示方法，这种虚假的不同能够导致错误的结论。在一个针对电话呼叫详细信息进行分析的项目中，每一个要研究的市场都以一种不同的方式表示一个检查自己语音邮件的呼叫。在第一个城市中，从与语音邮箱相关联的电话呼叫该邮箱，呼叫被记录为主叫电话和被叫电话相同；而在另外一个城市，相同的情况则可能通过把特定的不存

在的号码作为被叫电话来完成；在第三个城市中，则记录拨打语音邮件的实际电话号码。了解这些不同城市间语音邮件处理习惯的不同之处，就要求我们以通用格式来体现这些数据。

相同的数据集合包含对一些州的多种缩写形式，在某些情况下，某个特定的城市需要同本州的其他城市分开计算。如果这样的问题没有得到解决，我们就会发现自己建立的打往加州的呼叫模式模型不包含打往洛杉矶的数据（洛杉矶是美国加州南部的一个城市）。

3.4.6 第六步：变换数据，获取信息

数据组合起来后，主要的数据问题也得到了修复，仍然需要准备数据以便分析，包括添加一些字段以表达某些信息，也可能会涉及消除离群值、数值变量的装入、对不同类别变量的分组归类、使用对数函数（logarithms）进行变换以将某些计数转换为比例，等等。数据准备是一个非常重要的主题，我们的同事 Dorian Pyle 已经写了一本这方面的书 *Data preparation for Data Mining*（Morgan Kaufmann，1999），这几乎是每个数据挖掘者案头必备的书。上述问题将在本书的第 17 章进行讨论，这里给出一些这类转换的例子。

1. 捕捉趋势

大多数公司的数据含有时间序列，如营业额信息、使用情况、合同之类的每月快照。绝大多数的数据挖掘算法对时间序列数据不进行处理。相互独立的对单个月份的数据调查不能发现诸如“三个月收入下滑”这样的信号。数据挖掘者能够通过添加一些衍生变量来获取趋势信息，例如通过最近一个月的支出与前一个月支出的比率（ratio）可以获取短期趋势，最近一个月的支出与去年同期的比率则可以获得一个长期趋势（long-term trend）。

2. 创建比率及变量的其他组合

趋势分析仅仅是通过合并多个变量来获取信息的一个例子，还有很多其他例子。这些附加字段通常可以通过已有的字段导出，有经验的分析师对导出方法应该是很清楚的，但是仅仅通过软件却不容易考虑周全。典型的例子包括：

```
obesity_index = height²/weight
PE = price/earnings
pop_density = population/area
rpm = revenue_passengers * miles
```

把该领域专家认为重要的能够表达关系的字段添加进来，能够使得专家意见在挖掘过程中起到一定的作用。

3. 将计数转化为比率

许多数据集都包含计数（count）或者金额，这些计数和金额本身并无特殊意义，它们也会随其他的值而变化。大家庭会比那些相对较小的家庭在食品杂货上花费更多，他们会在农产品、肉制品、袋装商品、清洁产品和其他任何东西上花费更多。比较不同家庭在任意一个类别的消费金额，比如面包店，得到的结果都是大的家庭花销更多。比较每个家庭在每个类别中花销的比率应该是更有意义的。

通过比较基于纽约州的城镇数据集的两副图，可以将计数转换为比率的值看得很清楚。图 3-9 比较了通过劣质管道与流行的木材取暖的房屋的数量。它们之间的关系可见，但是对比并不明显。在图 3-10 中，劣质管道的房屋数量转化成劣质管道的房屋比率，关系就非常明显了。那些有很多劣质管道房屋的城镇也有许多使用木材取暖的房屋。这是否意味着燃烧木材产生的烟破坏了管道呢？重要的是要记住，我们发现的仅仅是模式的相关性，而不是因果问题。

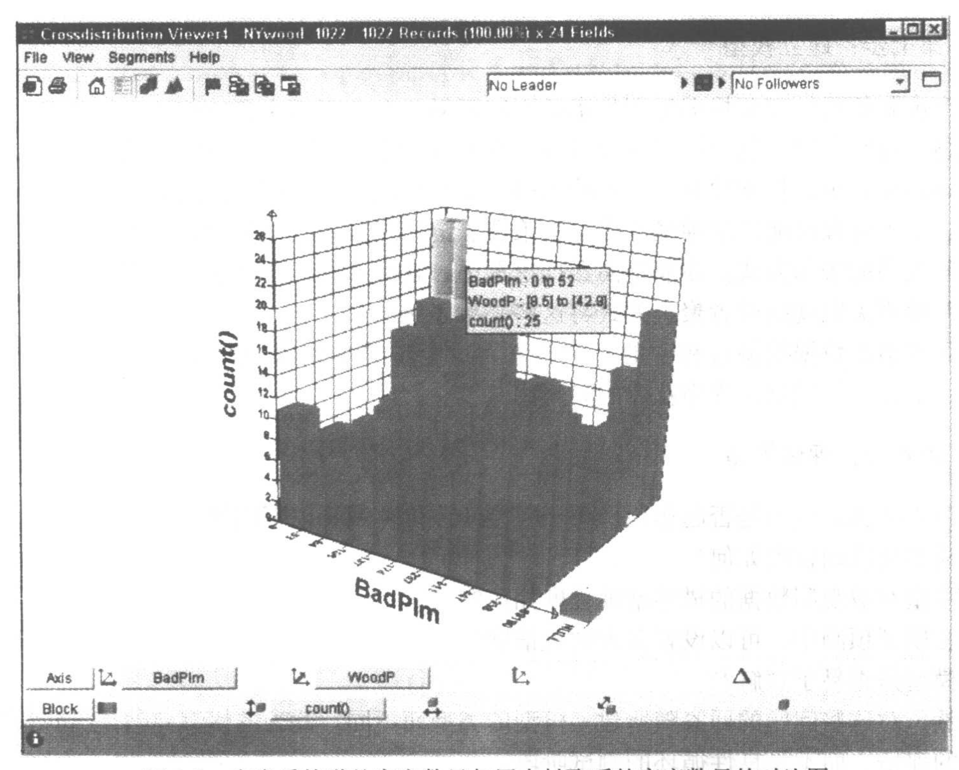


图 3-9 有劣质管道的家庭数目与用木材取暖的家庭数目的对比图

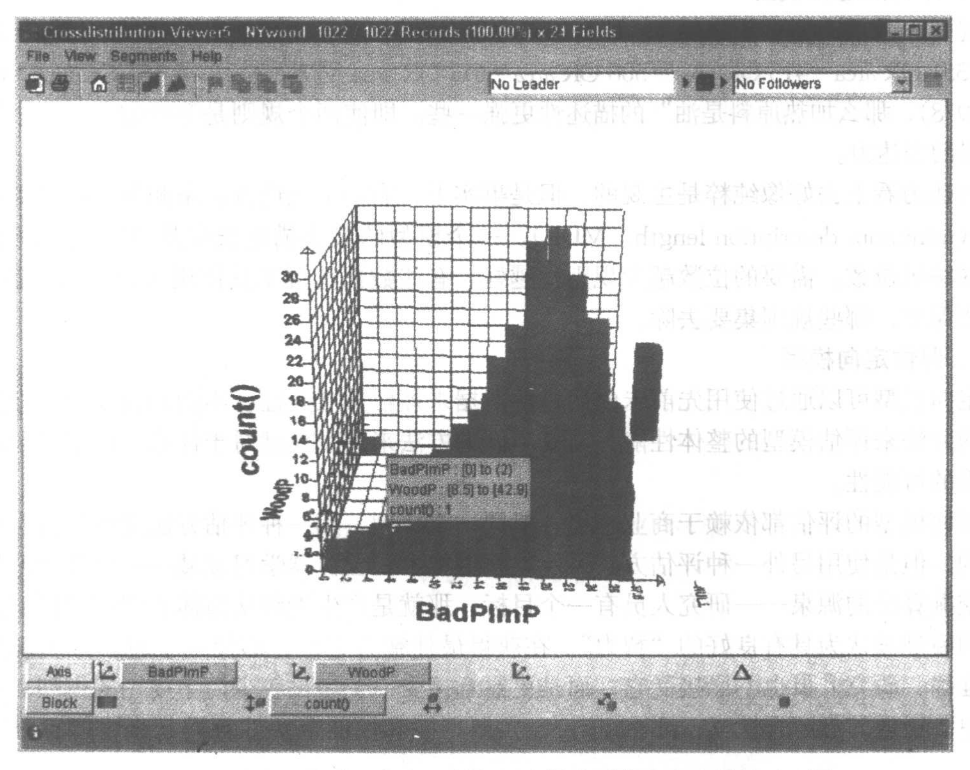


图 3-10 有劣质管道的家庭比例与用木材取暖的家庭比例的对比图

3.4.7 第七步：建立模型

这个步骤的具体细节根据使用技术的不同而异，在关于每个具体挖掘方法的章节中有相应的描述。通常来说，创建模型的大多数工作都在这个步骤进行。在定向数据挖掘（directed data mining）中，根据独立的或者输入的变量，训练集用于产生对独立的或者目标的变量的解释。这个解释可能采用神经网络、决策树、链接图或者其他表示数据库中的目标和其他字段之间关系的表示方式。在非定向数据挖掘中，就没有目标变量了。模型发现记录之间的关系，并使用关联规则或者聚类方式将这些关系表达出来。

建立模型是数据挖掘过程中的一个步骤，这个步骤由数据挖掘软件自动实现。正因如此，这一步在数据挖掘工程中花费的时间相对较少。

3.4.8 第八步：评估模型

这一步将决定模型是否起作用。关于模型的评价要回答以下问题：

- 模型的准确程度如何？
- 模型对被观测数据的描述精确程度如何？
- 在模型预测中，可以设置多大的置信度？
- 模型是否易于理解？

当然，对这些问题的回答随所建立模型的类型而不同。此处的模型评估是指对模型的技术优势的评估，而不是良性循环的测试阶段。

1. 评估描述性模型

规则："如果（state='MA'），那么加热原料是油"，看上去比规则："如果（area=339 OR area=351 OR area=413 OR area=508 OR area=617 OR area=774 OR area=781 OR area=857 OR area=978），那么加热原料是油"的描述性更强一些。即使两个规则是等价的，第一个似乎具有更强的表达力。

表达力看上去好像纯粹是主观的，但是事实上，有一个理论方法来测量——即最小描述长度（minimum description length，MDL），一个模型的最小描述长度是规则及其所有例外列表的编码位数。需要的位数越少规则就越好。有些数据挖掘工具使用MDL来决定哪些规则集要保留，哪些规则集要去除。

2. 评估定向模型

定向模型可以通过使用先前未使用过的数据来评估其准确性。不同的数据挖掘任务需要不同的方法来评估模型的整体性能，需要不同的方法来判断模型对于任意的特定记录产生准确结果的可能性。

任何模型的评估都依赖于商业环境，对同一个模型，用一种评估方法进行评估时可能是很好的，但是使用另外一种评估方法可能就很糟糕了。在机器学习领域——机器学习是许多数据挖掘算法的源泉——研究人员有一个目标，那就是产生能够从整体上理解的模型。易于理解的模型被认为具有良好的"智力"。在获得最佳智力方面，即使一个包含许多规则的模型更准确，研究人员也不愿接受它，而是更愿意接受一个包含较少简单规则的模型。在商业环境中，这样的可解释性（explicability），大概不能和性能重要性等同起来，但有时可能更重要。

模型评估可以在整个模型层次或者在单个预测层次上进行。在整个模型层次上具有相同准确度的两个模型，可能在单一预测层次会具有相当不同的水平。以决策树为例，它会有一个整体的分类差错率，同时对于其每个分支和枝叶也都会有一个差错率。

评估分类器（classifier）和预测器（predictor）

对于分类和预测任务，准确度（accuracy）是根据差错率来测量的，差错率是指被误分类的记录的百分比。在对新的记录数据进行分类时，基于预分类测试集的分类差错率可以用作对期望差错率的一个估计。当然，这个过程也只在测试集能够代表大量数据的普遍特征时才是有效的。

要确定模型的差错率，我们推荐的方法是使用测试数据集来测量，这个测试数据集应该是从与训练集和验证集相同的大量样本集合中抽取，但是要和它们分离开来。在理想的情况下，这种测试集的数据应该比模型集中的数据更新，然而在实际工作中，这通常是不可能的。

将差错率作为评价工具的问题在于差错的程度是不同的，有些差错可能比其他差错严重得多。医学界有一个熟悉的例子可以说明这个问题，对严重疾病检查中，错误的阴性结果能够导致病人耽误治疗，最终可能危及生命；而一个错误的阳性结果充其量也不过是让病人进行进一步的检查（当然有可能花销较大，或者具有侵害性）。如图3-11所示，可以使用含混矩阵（confusion matrix）或者正确分类矩阵（correct classification matrix）区分错误肯定和错误否定。有些数据挖掘工具允许将每一种类型的误分类与差错成本联系起来，这样建立的模型就可以使差错成本最小化，而不是将误分类率最小化。

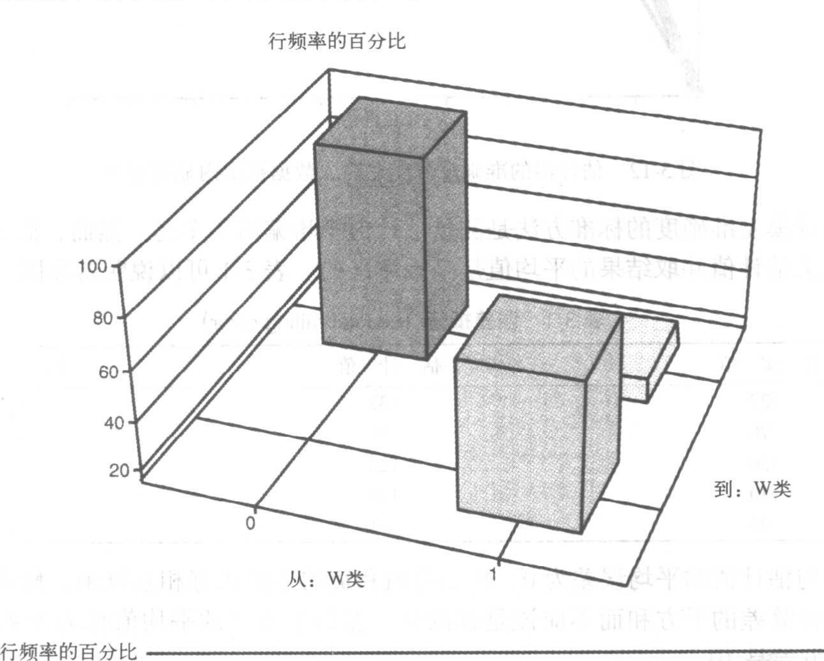


图3-11 含混矩阵交叉表预测的结果与实际结果比较

评估估计器（estimator）

对于估计任务，准确度是根据预测得分与实际测量结果的差异来描述的。任意一个估计的准确度和模型整体的准确度都是令人感兴趣的。一个模型可能对于某些输入值的范围非常准确，而对于另一些可能就非常不准确。图 3-12 是一个线性模型，这个模型基于一个产品单价来估计总收益。这个简单的模型在一个价格范围内运行得相当好，但是当价格达到产品需求弹性（销售量的变化率与价格的变化率的比值）大于 1 的水平时，其表现就非常糟糕了。弹性大于 1 意味着价格的进一步增长将会导致总收益的下降，原因在于每单位增加的收益被销售数量的下降抵消了。

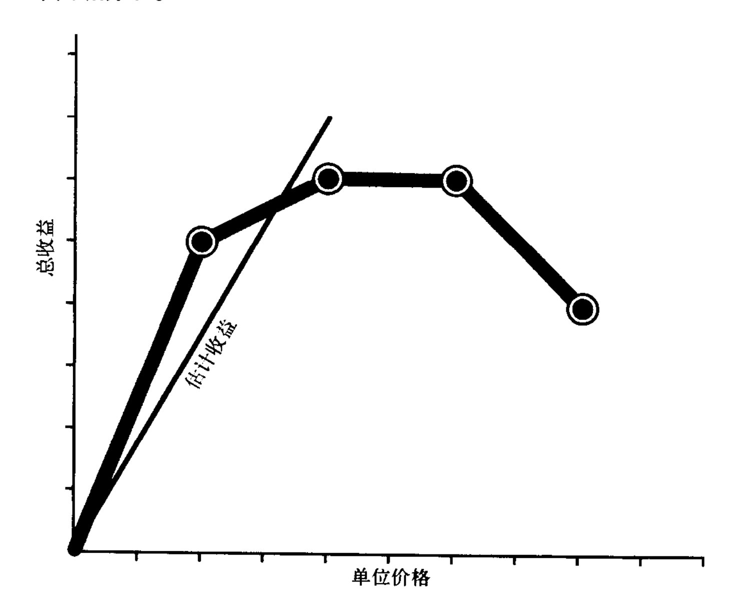


图 3-12 估计器的准确度可能在输入数据范围内显著变化

描述估计模型准确度的标准方法是测量估计值平均偏离了多远。然而，简单地从每个点的真实值减去估计值并取结果的平均值是毫无意义的。表 3-1 可以说明其原因。

表 3-1 误差抵消（countervailing error）

真 实 值	估 计 值	误 差
127	132	−5
78	76	2
120	122	−2
130	129	1
95	91	4

真实值与估计值的平均误差为 0，正的差值和负的差值正好相互抵消。解决这个问题常用的办法是计算差的平方和而不应该是差的和。差的平方和的平均值称为*方差*（variance）。表 3-1 中的方差是 10。

$$(-5^2+2^2+(-2)^2+1^2+4^2)/5=(25+4+4+1+16)/5=50/5=10$$

方差越小，估计就越准确。方差的一个缺点是与估计本身的单位不统一。用美元做单位估计价格时，估计偏离值用美元做单位比用美元的*平方*做单位要有用得多，所以，通常我们

会取方差的平方根作为度量，并称之为*标准差*（standard deviation）。表 3-1 中的标准差为 10 的平方根，即 3.16 左右。对于我们的目的来说，只要知道标准差能够测量估计值与真实值究竟偏离多远就可以了。

3. 利用提升度来比较模型

无论是使用神经网络、决策树、遗传算法还是其他方法，定向模型都是为完成某个任务创建的。为什么不评价一下它们在分类、估计与预测方面的能力呢？比较分类模型性能的最常用方法是使用称为*提升度*（lift）的比率。这个度量也能够用于比较为其他任务设计的模型。提升度实际上测量的是，当模型用于从总体中选择一个群组时，这个指定群组集中度的变化量。

$$lift = P(class_t | sample)/P(class_t | population)$$

可以举个例子来说明这个问题。假定我们建立了一个模型，用于预测哪些人比较容易对直接的邮件诱惑做出响应。通常，我们使用预分类训练数据集建立模型，如果需要的话，也会使用预分类验证集进行验证。现在，我们可以使用测试集计算模型的提升度。

分类器对测试集中的记录标记“预测响应”或者“预测不响应”。当然，这不是每次都会给出正确结果，但是如果模型足够好的话，标记为“预测响应”的记录组，实际响应者的比例会比测试集整体实际响应者的比例更高。考虑这样一个结果：如果测试集包含 5%的实际响应者，而样本中包含 50%的实际响应者，模型给出的提升度就是 10（50 除以 5）。

能够得到最高提升度的模型一定是最佳模型吗？有半数响应者的列表当然会比另一个只有 1/4 响应者的更好，对吗？不一定，如果第一个列表中仅仅只有 10 个人的话，就不是最好的！

关键在于提升度是样本大小的函数。如果分类器只选出了 10 个可能的响应者，这时正确率为 100%，提升度会达到 20——当总体中只有 5%的响应者时，这是可能的最高值。把用于将某人分为可能响应者的置信度水平放宽的话，邮件列表将会变得很长，提升度就会下降。

在使用数据挖掘工具时，如图 3-13 所示的图表是非常常见的。根据模型预测的响应情况，对所有潜在客户排序就能生成这些图表。当邮件列表增加的时候，我们沿着这个列表走得越来越远。X 轴表示的是人群总体中收到邮件者的百分比，Y 轴表示的是我们能联系到的所有响应者的百分比。

如果没有使用模型，给人群总体中 10%的人发送邮件可能会接触到 10%的响应者，发送 50%则会接触到 50%的响应者，给每个人发送的话会接触到所有的响应者。这种群体发送邮件的方法可以用图中向上倾斜的直线表示。另一条曲线表示的是，如果用模型选择邮件的接收者情况会如何。通过使用模型发现，只要给占总人口 10%的人发邮件，就会得到 20%的邮件响应者，所以只要对总人口的一半人发邮件加以诱惑，就可以接触到 70%以上的响应者。

如图 3-13 所示的图表常被称为*提升度图*（lift chart），尽管实际上画出的是*累积响应*（cumulative response）或者称作*集中度*（concentration）。图 3-13 表示的是对应于图 3-14 中响应图的实际提升度图。这个图清楚地表明，随着目标列表规模的增加，提升度呈下降趋势。

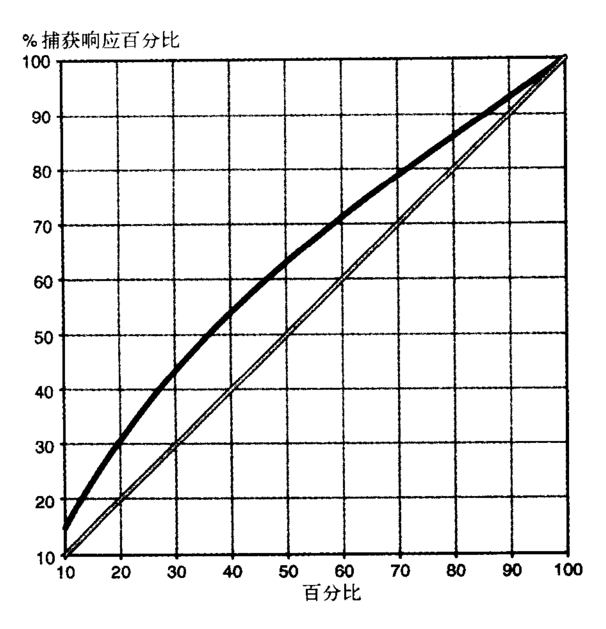


图 3-13 目标邮寄与大众邮寄的累积响应的比较

提升度存在的问题

提升度解决了如何比较不同模型性能的问题，但是还不能回答最重要的问题：一个模型是否值得花时间、精力和金钱来创建？给提升度是 3 的客户群体邮寄会是有利可图的活动吗？

为了将花销与收益纳入到考虑范围，如果没有业务的更多背景知识，是无法回答这些问题的。另外，当两个模型被用于同样的或者相近的数据时，提升度也是一个非常便利的比较性能的工具。要注意的是，当测试集的输出结果有相同密集度时，两个模型的性能只能使用提升度来比较。

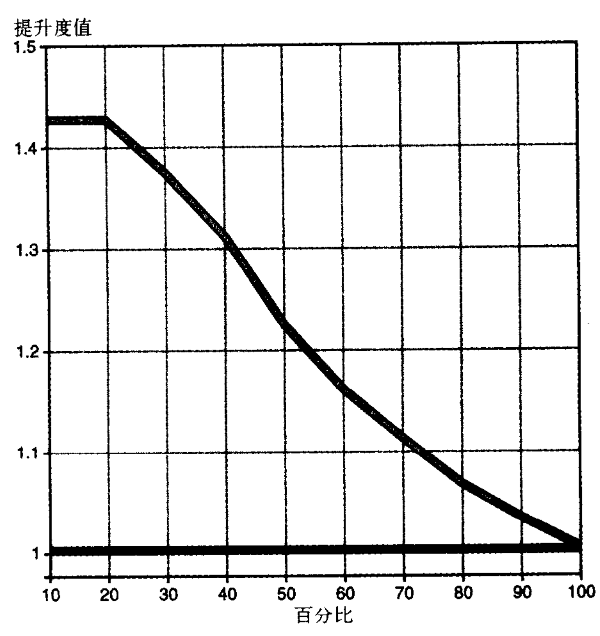


图 3-14 提升度图，初始值很高然后逐渐降为 1

3.4.9 第九步：部署模型

部署模型（deploy model）就是将其从数据挖掘的环境转移到评分环境。这个过程可能很简单，也可能很难。在最坏的情况下（这种情况我们不止在一个公司遇见过），模型是用其他地方都不能够使用的软件在特殊环境中开发出来的。为了调用这个模型，程序员需要把模型的描述打印出来，然后用另外一种编程语言将其重新编程，以便在评分平台上运行。

更常见的一个问题是模型使用了未在初始数据中出现的输入变量。由于模型的输入至少源于字段，而这些字段最初是从模型集中提取出来的，所以这应该不是什么问题。不幸的是，数据挖掘者对数据进行变换，但并不总是能够保持变换后的记录“干净”和可重用。

部署数据挖掘模型的挑战在于，它们经常被用于为大型的数据集评分。在某些环境下，每天都要为上百万个客户记录中的每一条记录更新行为得分。评分值仅是数据库表中的一个附加字段，表示的经常是一种可能性或者是倾向性，因此通常取介于 0 到 1 之间的数值，当然，也并非一定如此。评分值也可能是聚类模型给出的类标记（class label），或者是带有可能性（probability）的类标记。

3.4.10 第十步：评估结果

图 3-14 中的响应图比较了使用或者不使用预言性模型时，在给定邮资额度情况下的响应者数量。一个更有用的图应该能够展示在市场营销活动中，对于给定的花销能够带来多少利润。毕竟，如果开发这个模型很昂贵的话，大众邮寄比目标邮寄的单位成本可能更划算。

- 建立市场活动和支持模式的固定成本是多少?
- 对每一个服务接受者而言的花费是多少?
- 对每一个执行服务的响应者而言成本是多少?
- 一个积极响应的价值是多少?

将这些数据填入一个电子数据表中，使其能够以美元来测试模型的影响。累积响应图能够变成累积利润图，这个图决定了分类邮件列表应该从哪个位置截断。例如，如果开展市场营销活动的固定价格很高，且对每一个接受者的优惠服务价格也相当高（如无线电话公司通过赠送手机或者抹掉更新费用来换取客户的忠诚），公司会因为追求极少数潜在客户而赔钱，因为这种计划的响应者数量不够多，不足以弥补实行该计划所需要的固定成本。另一方面，如果要对太多人提供优惠，又要花费很高的可变成本，也会损害公司利益。

当然，利润模型成功与否取决于输入值的好坏。即使在固定成本和可变成本很容易计算的情况下，每个响应者的预言性价值也是很难估计的。断定一位客户的价值有多大的过程超出了本书的范围，但是一个良好的估计有助于测试数据挖掘模型的真正价值。

最后，最重要的度量是投资回报率。在测试集上测量提升度有助于选择正确的模型。基于提升度的收益率模型有助于决定如何使用模型给出的结果，但在本领域内进行测量也是非常重要的。在数据库市场应用中，这需要取消对照组，根据不同的模型得分仔细地追踪客户的响应状况。

3.9.11 第十一步：重新开始

每一个数据挖掘项目引出的问题都比它回答的问题要多。这是一件好事，意味着以前不

可见的新关系，现在已经变得可见。新发现的关系提出用于测试的新假定，数据挖掘过程再一次从头开始。

3.5 小结

数据挖掘来自两种形式。定向数据挖掘就是通过对历史数据的查找，找到能够解释特定输出结果的模式。定向数据挖掘包括分类、估计、预测和建立简档等任务。非定向数据挖掘通过同样的记录搜索，发现令人感兴趣的模式，包括聚类、发现关联规则和描述等任务。

数据挖掘缩短了商业与数据的距离。就这一点而言，假设测试是这个过程中非常重要的部分。然而，从本章所讲的主要内容可以看出，对于那些粗心大意的人来说，数据挖掘充满陷阱，遵循基于经验的方法论则能够帮助避开这些陷阱。

第一个关键问题是将相关的商业问题转换为数据挖掘可以解决的六个任务之一：分类、估计、预测、关联分组、聚类和建立简档。

第二个问题在于找到可以转化为可操作信息的合适数据。一旦找到了合适的数据，就要进行彻底的研究。探究的过程很容易揭示数据存在的问题，同时也有助于数据挖掘者对数据的直观理解。下一步就是创建模型集，并将其划分为训练集、验证集和测试集。

数据转化对于以下两个目的是必要的：1）修复数据存在的问题，如缺失值、取很多值的分类变量等；2）通过创建表示趋势和其他比率与组合的变量，将信息显示出来。

一旦数据准备好了，创建模型就是相对容易的过程。每一种类型的模型，都有各自用于评估的度量。但是也存在独立于模型类型的评估工具，其中最重要的两个工具是：1）提升度图，能够显示被评估模型如何提高目标变量期望值的集中度；2）含混矩阵，显示每一个目标类的误分类差错率。在下一章中，将利用几个来自实际数据挖掘项目的案例展示实际操作的方法论。

第4章 数据挖掘在市场营销和客户关系管理中的应用

一部分人从技术前景方面发现数据挖掘技术是令人感兴趣的，然而大多数人最终是把这项技术作为一种手段而对它感兴趣。这种技术不是存在于真空中，而是存在于整个业务活动的过程中。本章要讲述的内容就是与业务活动过程相关的。

本章内容围绕能够在数据挖掘中用到的一套业务目标。每一个选定的业务目标都与可以解决选定问题的特定数据挖掘技术相联系。本章中所选择的业务主题大致按照客户关系的复杂性由浅入深加以表述，从与知之甚少的潜在客户的沟通问题，逐渐转到可能涉及多种产品、多种通信渠道和其他日益个性化互动的现有客户关系中所呈现出的各种各样的数据挖掘领域。

下面在讨论业务应用过程的同时，会适当介绍有关技术资料，但数据挖掘技术的具体细节将在以后的章节中介绍。

4.1 寻找潜在客户

寻找潜在客户（prospecting）这个英文单词似乎是一个开始讨论数据挖掘商业应用的好起点。英文中“prospect”作为动词的最初定义来自传统的采矿业，意思是*探寻矿物或石油*。作为一个名词，“prospect”可解释为具有可能性的、能引起开采油田或者挖掘矿产的联想的那些事物。在市场营销方面，prospect 指那些通过正确方式接近有可能成为客户的某个人，即潜在客户。不论作为名词还是动词，在使用数据挖掘技术以确定未来谁将会成为有价值的客户这一商业目标上，它所代表的意义是相通的。

对于大多数交易，地球上超过六十亿的人口中只有相对极少的一部分会是实际的潜在客户，绝大多数人因地理、年龄、支付能力和对产品或服务的需求等各种原因而被排除在外。例如，提供家庭抵押贷款的银行自然会严格控制投递范围，把这类促销邮件寄给居住在这家银行授权经营区域内的那些住户；卖庭院秋千的公司，喜欢把目录寄给从地址上看起来可能有庭院、有小孩的家庭；杂志要瞄准具有相应语言阅读能力并且对登广告感兴趣的那些读者，诸如此类的例子还有很多。

数据挖掘能在探查潜在客户方面扮演多种角色（role），其中最重要的是：

- 识别好的潜在客户
- 为接近潜在客户选择沟通渠道（communication channel）
- 针对不同的潜在客户群，选择合适的信息

尽管所有这些都很重要，但第一项——识别好的潜在客户——应该是数据挖掘最广泛应用的一个方面。

4.1.1 识别好的潜在客户

多数公司对“好的潜在客户”（good prospect）的最简单定义是：对可能成为客户至少会表现出一点兴趣的某个人。更复杂的定义就需要进一步斟酌。确实，好的潜在客户不仅要

对成为客户感兴趣，他们还必须能买得起商品，成为客户对公司是有利可图的，并且不太可能欺骗公司而且会及时支付账单。而且，如果善待他们，他们将成为忠实的客户并推荐另外的客户。不论潜在客户的定义多简单或者多复杂，首要的任务是要找准他们。

不管信息是通过广告传送还是通过更直接的渠道，如邮寄、电话或者电子邮件，目标明确（targeting）是重要的。即便是广告牌上的信息，在一定程度上也是有针对性的：在通向机场的公路边容易发现航空公司和汽车出租公司的广告牌，因为使用这些服务的人就在那些驾车路过的人群之中。

要把数据挖掘用于这一问题，首先要定义具有什么特征的人是好的潜在客户，然后找出能够瞄准具备那些特征的人们的方法。对于许多公司来说，要使用数据挖掘识别好的潜在客户，第一步是建立响应模型。在本章中稍后是关于响应模型的详细讨论，阐述利用它们的各种方法，以及它们能做什么、不能做什么等。

4.1.2 选择沟通渠道

寻找潜在客户需要沟通。一般来说，公司总是试图以几种方式与潜在客户沟通。一种方式是借助公共关系，即是指鼓励媒体专题报道公司事务，以及以口头方式传播公司积极的信息。虽然公共关系对于某些公司（如 Starbucks 和 Tupperware）很有效，但不是定向的市场营销渠道。

我们更感兴趣的是广告和定向市场营销。广告可以做在任何地方，从火柴盒到一些商业网站上弹出的令人讨厌的窗口，从重大体育赛事的电视直播，再到电影中的物品布置。从这个方面看，广告针对的是具有共同特性的人群，然而广告并不能针对某个个体给出个性化信息。接下来的部分讨论了通过匹配地理区域档案和潜在客户档案，以选择正确的广告场所。

定向市场营销确实允许定制个性化信息，通常的方法是打电话、发电子邮件、寄明信片或邮寄五光十色的彩色目录。本章的稍后部分是关于差别响应分析的，解释了数据挖掘如何帮助决定哪一种沟通渠道对哪一组潜在客户是有效的。

4.1.3 遴选适当的信息

即使在销售相同的日用产品或服务时，对不同的人也要适当地提供不同信息。举例来说，同一张报纸，吸引某些读者的可能主要是运动版面，而对其他人则可能主要是政治和艺术版面。当产品本身存在许多不同品种，或者有多种产品可供销售时，选出正确的信息就更重要了。

即使对于单一产品，正确的信息仍是重要的，一个经典的例子是价格和便利之间的权衡。一些人对价格很敏感，愿意到大商场购物，喜欢在深夜里打电话，总是愿意转乘飞机，并且把星期六晚上安排到行程中去；而另外一些人则可能愿意为更方便的服务支付额外费用。价格信息不仅可能无法刺激追求便利的人，而且可能冒险：当客户愿意支付更多钱的时候，却将他们引向了利润更少的产品。

本章描述了如何将简单的、单一促销活动响应模型组合起来，创建促销活动与客户相匹配的最佳后续服务模型。协作过滤（collaborative filtering）也是一种有用方法，它把具有相似意向的客户进行分组，而这些客户组对相同的产品服务可能做出相似响应，协作过滤方法将在第 8 章讨论。

4.2 为选择正确的广告场所进行数据挖掘

寻找潜在客户的一种方法是寻找与现有客户类似的人。举例来说，经过调查，一本全国性的出版物认定它的读者具有下列特征：

- 59%的读者受过高等教育；
- 46%属于专业技术或行政职位；
- 21%的家庭年收入超过75 000美元；
- 7%的家庭年收入超过100 000美元。

理解这一系列用户特征数据（以下简称为“简档”），在以下两个方面有助于该出版物的发行：首先，通过瞄准与该描述相匹配的潜在客户，可以增加自己营销工作的响应比率。其次，有了这些受过良好教育的、高收入的读者群，可以把出版物上的广告空间出售给那些希望其产品信息能够到达这些受众群体的公司。因为本部分的主题是以潜在客户为目标，让我们看看该出版物是如何利用这些简档来强化其客户发掘工作的。基本思路很简单，当出版物想要通过无线广播做广告的时候，应该寻找其听众与这些简档相匹配的广播电台。当它要在商店柜台上放随手可取的广告卡片时，应该把它们放在与这一简档相匹配客户的居住地附近的商店柜台上。如果想打电话推销，应该把电话打给那些与这一简档相匹配的人。因而数据挖掘面临的首要问题就是给“简档匹配”下一个好的定义。

4.2.1 谁匹配简档

决定一位客户是否匹配某简档的方法，是衡量客户和简档之间的相似性，我们称之为距离。有几种数据挖掘技术使用测量相似性距离这一概念。在第8章将要讨论的基于存储的推理（memory-based reasoning），就是一种把具有“相邻近属性”的已知记录归类的分类技术。自动聚类检测（第11章的主要内容）则是另一种数据挖掘技术，它通过计算两个记录之间的距离，查找记录之间彼此类似的簇。

对这一简档的例子，其目的很简单，只是定义一个距离度量，以决定潜在客户与该简档相匹配的程度。由测量结果组成的数据只是在某个特定时间内客户的一个快照。对于这一数据，哪种测算方法更有意义？特别是，对于以百分数方式（58%受过高等教育，7%收入年超过100 000美元）描述简档的事实如何处理？反之，对于“受过或者没有受过高等教育”、“年收入超过或者不超过100 000美元”的个体，这样的数据又该如何处理？

下面通过一个例子来说明这个问题。设想有两个调查参与者：Amy受过高等教育，每年赚80 000美元，是自由职业者；Bob中学毕业，每年赚50 000美元。哪一个与读者简档更接近？问题的答案取决于采用哪种方式对比。表4-1显示了仅仅使用该简档和简单差距度量产生得分的一个方法。

表4-1 通过与每个人口统计学度量比较来计算个体的匹配度得分

	读者比例	是得分	否得分	AMY	BOB	AMY得分	BOB得分
受高等教育	58%	0.58	0.42	是	否	0.58	0.42
专业或行政	46%	0.46	0.54	是	否	0.46	0.54
收入>75 000美元	21%	0.21	0.79	是	否	0.21	0.79

（续）

	读者比例	是得分	否得分	AMY	BOB	AMY得分	BOB得分
收入>100 000 美元	7%	0.07	0.93	否	否	0.93	0.93
总计						2.18	2.68

这个表按照匹配每一特征的受众比例计算得分。比如，“58%的读者受过高等教育”这一项，Amy 因具备这一特征得 0.58 分。Bob 没有大学毕业，得 0.42 分，因为其他 42%的读者假定是没有大学毕业。对于每一特征依此类推，最后把得分加起来：Amy 最终得 2.18 分，Bob 得到了更高的 2.68 分。他的高分表示他比 Amy 更匹配目前读者的简档。

这个计算方法的问题是，虽然从分值结果看 Bob 比 Amy 与简档更匹配，但 Amy 似乎更接近于出版物目标受众，即受过高等教育、有高收入的那些个体。很明显，通过把读者简档与将美国人口作为整体的人口统计学数据对比，从而锁定目标的方法是成功的。这就提示了一个人是否适应成为出版物受众的更可信的度量方法，即不仅仅要考虑读者特征，还要重视总体人群特征。通过与确定出版物受众相似的方法，可以测量潜在客户与总体人群的差异程度。

与总人口相比，读者群受过更好的教育，更专业，收入也更高。在表 4-2 中，“指数（index）”列是将具有特定属性读者的百分比除以具有该属性的人占总人口的百分比计算得出的，因为读者特征百分比对具有读者特征的人占总人口百分比有特别贡献。现在我们看到，在“受高等教育”这一项，读者受高等教育的比例差不多三倍于整个人口受高等教育的比例。类似地，读者未受高等教育的比例大约只相当于总人口比例的一半。通过用指数作为得分考察每一特征，Amy 得分 8.42（2.86+2.40+2.21+0.95），而 Bob 得分 3.02（0.53+0.67+0.87+0.95）。基于指数的得分与出版物目标受众匹配性要好很多。新的得分更有意义，因为考虑到了目标受众区别于美国人口整体的一些额外信息。

表 4-2　考虑在人口中所占比例计算得分

	读者比例	是占总人口	指数	读者比例	否占总人口	指数
受高等教育	58%	20.3%	2.86	42%	79.7%	0.53
专业或行政	46%	19.2%	2.40	54%	80.8%	0.67
收入>75 000 美元	21%	9.5%	2.21	79%	90.5%	0.87
收入>100 000 美元	7%	2.4%	2.92	93%	97.6%	0.95

提示：当比较客户简档的时候，记住把总人口简档考虑在内是很重要的。正是由于这一原因，使用指数得分往往比使用原始分值更好。

第 11 章描述了基于两种角度之差的相似性的有关概念。在该方法中，每一被测属性被认为是一个独立的坐标点。取每一属性的平均值作为原点，当前读者可用一个矢量描述，代表了他（或她）不同于整体的偏离值和方向。代表潜在客户的数据也是一个矢量。如果这两个矢量之间的夹角很小，则潜在客户与在同一方向上的群体有所不同。

4.2.2　测量读者群组的匹配度

基于指数得分（index-based score）的思想可以扩展到更大的群组。因为公司不一定掌握评价每一位客户或潜在客户所需要的总体统计特征的特定数据，所以这一点很重要。幸运

的是，前述特征都是可以通过美国人口普查来获得的人口统计学数据，并且可以按如人口普查区域（census tract）等地理分布方法度量（见“人口普查区域数据”部分）。

这里描述的过程，是为该出版物按照其匹配度来给每一人口普查区域评定等级，意图是评估每一人口普查区域符合该出版物读者特征的比例。举例来说，如果一个人口普查区域成年人群 58%受过高等教育，那么其中的每一个人在这一特征的匹配度得分为 1；如果 100%都受过高等教育，那得分仍是 1——完全符合我们能做到的最好程度。然而，如果只有 5.8%受过高等教育，那么这项特征的匹配度得分就是 0.1。最后总的匹配度得分是每一特征得分的平均值。

图 4-1 提供了曼哈顿三个人口普查区域的实例。每个区域都有所要考虑的四项特征的不同百分比，这些数据可以组合起来得到每一区域的总匹配度得分，这一得分代表了该区域人口匹配该简档的比例。需要说明的是，区域中的每个个体得分相同。

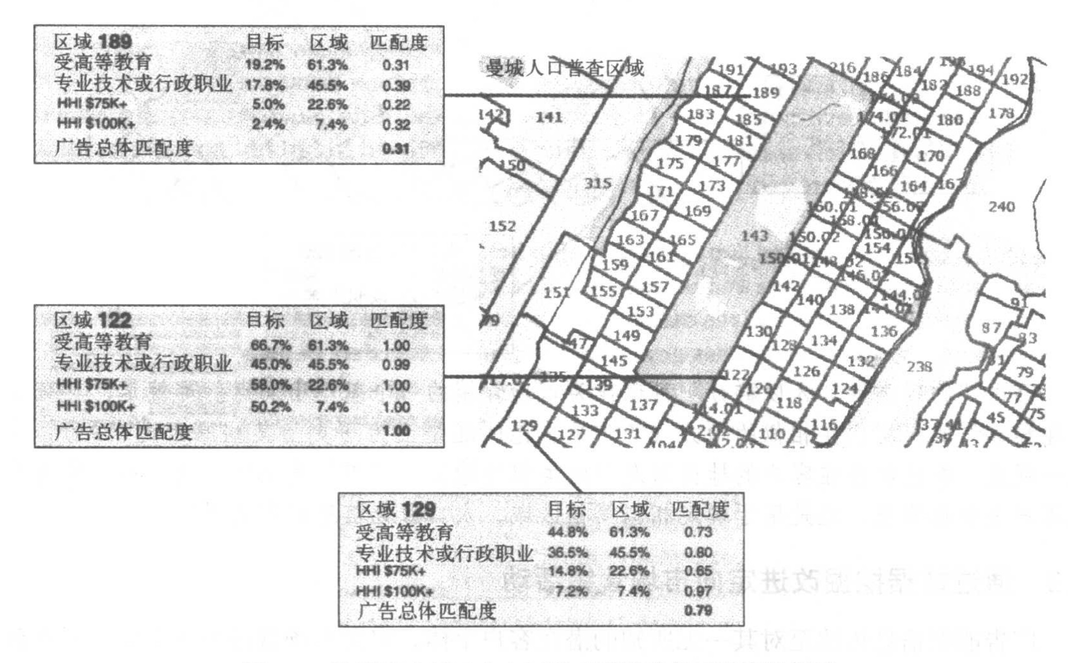


图 4-1 计算曼哈顿三个人口普查区域读者匹配度的例子

人口普查区域数据

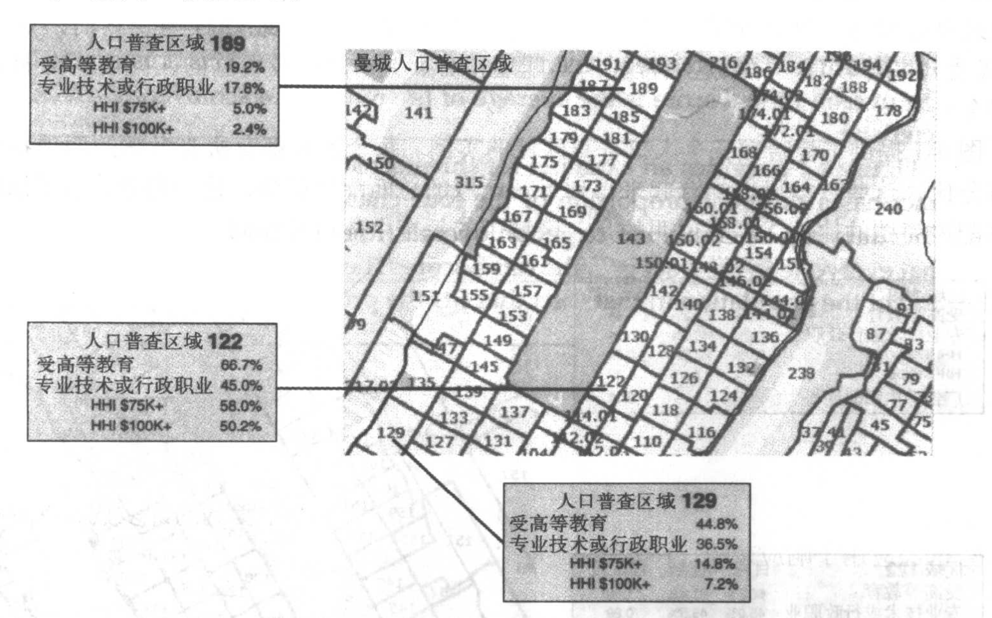
美国政府依法每 10 年进行一次人口普查，主要目的是分派每个州在众议院的席位。在满足这一要求的过程中，人口普查也提供了关于美国大众的很多资讯。

美国人口普查局（U.S.Census Bureau，www.census.gov）使用两种调查表调查美国大众，短表（short form）和长表（非特别目的调查表，如军事人员使用的）。大多数的人拿到短表，询问一些关于性别、年龄、种族和家庭成员等基本问题。大约 2%的民众得到的是长表（long form），问及关于收入、职业、通勤习惯、消费方式等一些详细问题，对这些调查表的响应提供了人口统计简档的基础。

人口普查局努力使这一信息在每十年进行的人口普查之间保持最新。人口普查局不发布被普查个体的信息，而是以小的地理区域为单元聚合这些信息。最常用的是人口普查区域，由大约 4 000 人组成。尽管人口普查区域大小各异，但它们比其他地理单元，如县和邮政区

划，在人口规模上更具有一致性。

人口普查确实有更小的地理单元，如街区和街区组合。然而，为了保护居民隐私，一些低于人口普查区域层次的数据是得不到的。从这些区划单元数据可以聚合得出县、州、都市统计区域（metropolitan statistical area，MSA）以及立法行政区等信息。下图显示了位于曼哈顿中心的一些人口普查区域。



谚语“物以类聚，人以群分”可以作为市场营销的一个基本出发点。那就是，具有相似志趣和喜好的人居住在相似的地方（不论是自愿的还是因为历史上某些原因造成的）。按照这一观点，在已经存在客户的地区以及与此类似的地区开拓市场是个好办法。不论是用于发现客户集中在哪里，还是用于确定相似简档区域，人口普查信息都有利用价值。

4.3 通过数据挖掘改进定向市场营销活动

广告能把信息传达至对其一无所知的潜在客户个体，定向市场营销至少需要一丁点额外信息，如姓名、通信地址、电话号码或电子邮箱等。信息越多，数据挖掘发挥作用的机会越多。最基本的，数据挖掘可以通过选择联系谁以改善目标定位。

实际上，起始层次的目标搜索并不需要数据挖掘，需要的只是数据。在美国有关人口整体的相当多数据是可以得到的。（很多其他国家可能稍微少一点），在许多国家，有一些公司汇编和出售各种各样家庭层次的数据，包括收入、孩子个数、教育水平，甚至业余爱好等。这些数据有一些是从公众档案中收集来的：家庭采购、婚姻、生育、死亡事件都是可以从县法院和行政登记机关获悉的公开记录，其他数据可以从产品登记表中收集，有一些是使用模型得来的。各个国家用于市场目的的数据使用管理条例不尽相同。在一些国家，数据可以按地址，但不能按姓名出售，另一些国家，数据只允许用于某些特定准许的目的。还有一些国家，数据几乎可以没有限制地使用，但只覆盖了数量有限的家庭。在美国，一些医疗记录类数据是完全禁止使用的，而对于信用历史之类的数据只能用于特定的经核准的目的，其余大部分则是不受限制的。

警告： 美国不论在可用家庭数据的商业范围还是用途上，其相对较少的限制都是与众不同的。尽管家庭数据在许多国家是可用的，但控制其用途的条例则各不相同，另外还可能有特别严格的条例控制个人数据的境外传输。在打算将家庭数据用于市场营销前，必须首先考察它们在市场推广方面是否可用，在使用这些数据时有什么法律限制，等等。

在诸如收入、是否有汽车或者是否有孩子等情况的基础上，家庭层次的数据（household-level data）可以初步整理，划分群体后直接使用。但问题是，即使采取了显著的筛选措施，与可能响应的潜在客户相比，剩余群的数目仍然很大。这样，针对潜在客户的数据挖掘应用的首要问题是确定目标——发现最可能对优惠服务做出实际响应的潜在客户。

4.3.1 响应建模

通常，定向市场营销活动的响应率一般是一位数。响应模型通过识别潜在客户，即谁更可能对直接诱导做出响应，来提高响应率。最有用的响应模型应该提供对可能响应的真实估计，但这不是必要条件，任何可以把潜在客户按响应可能性分级的模型都可以满足需要。给出一个分级列表后，直接面向市场的营销人员可以给列在表顶端的人们发邮件或打电话，以增加活动可触及的响应者的百分比。

下面的小节将描述可以利用模型得分促进定向市场营销的几种方法。下面的讨论内容与用以生成得分的数据挖掘技术无关，然而需要指出的是，本书中提到的许多数据挖掘技术能够并且已经应用于响应建模（response modeling）。

依据定向市场营销联合会（Direct Marketing Association，一个行业组织）的统计，100 000件的普通邮寄花费大约 100 000 美元，尽管价格可能因邮寄过程的复杂性而有所不同。这其中的一些费用，像开发创新性内容、准备美术品和印刷的初始要求，与寄件数量无关，剩余的费用则直接随邮寄数量不同而改变。已有的订单响应者名录或者订阅杂志人名录，可以每千人多少钱的价格购得，相关邮政用品费用和邮资也可以按相似基数计算。邮寄量越大，固定成本比例将变得越小，在总成本计算中变得越不重要。所以为简化计算，本书中的例子假定利用直接邮寄活动送达一个人要花 1 美元。尽管简单的邮寄成本少些，而很精美的邮寄成本会多些，但这种近似估计仍然是合理的。

4.3.2 优化固定预算的响应率

使用模型得分的最简单方式是用它们来列出等级。一旦潜在客户被指定了响应倾向（propensity-to-respond）得分，潜在客户列表就可以进行排序，把那些最可能响应的人排在列表的顶端，最不可能响应的人排在底部。许多建模技术能够用来生成响应得分，包括回归模型、决策树和神经网络等。

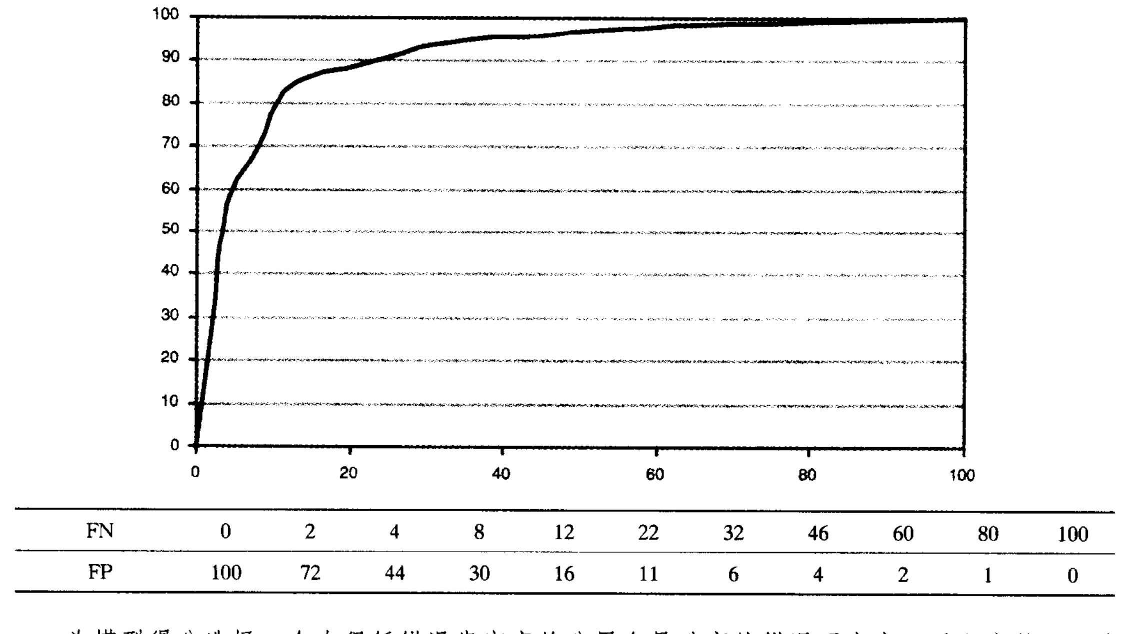
在没有时间或足够预算来送达所有潜在客户的情况下，列表排序就显示了它的意义。如果一些人必须被抛弃，略去那些不大可能响应的人就是有意义的。不是所有的交易都有必要略去一些潜在客户。市话公司可能认为城里的每个家庭都是潜在客户，并且它有能力一年内给每个家庭写几次信或打电话。因而当市场计划要求对每一个潜在客户提供相同服务的时候，没有多大必要进行响应建模！然而，对于为客户选择出适当的信息，以及预计潜在客户在多大程度上可能成为客户这样一些问题，数据挖掘仍然是非常有用的。

更可能的情形是，市场营销预算不允许被每一个潜在客户以相同水平占用。设想一个公司在其潜在客户列表上有100万个姓名，要花费300 000美元搞市场营销活动，每联络一个客户花费1美元成本。这个公司，我们称它为简单假设公司（the Simplifying Assumptions Corporation，SAC），可以用一个响应模型评定潜在客户列表得分，花300 000美元给列表中得分最高的300 000名潜在客户发送服务信息，能使得响应数量最大。这一行动的结果如图4-2所示。

ROC 曲线

模型用来产生得分。当分界得分（cutoff score）用来决定在营销活动中应该包括哪些客户的时候，客户实际上被分为两类——有可能响应，和不可能响应。评估分类规则的方法是察看它的差错率。在一个二元分类任务（只有两种可能的事件）中，总误分类率（misclassification rate）有两个分量——错误肯定率（false positive rate，FP）和错误否定率（false negative rate，FN），变更分界得分会改变这两种差错类型的比例。对于一个高分值意味着高响应可能的响应模型，选择一个高分作为界限意味着很少会有肯定错误（被标记为响应者而没有响应的人们），但会有更多的否定错误（标记为不响应但做出响应的人们）。

用ROC曲线可以表示实验测试时分界得分变化所导致的错误肯定率和错误否定率的变化关系。字母ROC代表接收器作业特性（Receiver Operating Characteristics），这个名字可以追溯到第二次世界大战，它的提出最初是用于评价雷达操作员正确识别雷达显示器上的点是敌舰还是无害航行物的能力。今天，ROC曲线更多是被医学研究者用来评估医学检测结果。错误肯定率作为X轴，用1减去错误否定率作为Y轴。下图中的ROC曲线反映了下面表格给出的一个错误情况测试。



FN	0	2	4	8	12	22	32	46	60	80	100
FP	100	72	44	30	16	11	6	4	2	1	0

为模型得分选择一个有很低错误肯定率的分界会导致高的错误否定率，反之亦然。一个好的模型（或医学检查）应该有一些得分能够分辨结果，并因此减少这两种类型的错误。在能够做到这一点时，ROC曲线向左上角凸起。ROC曲线下方的区域是该模型区分两个结果

的能力的度量，这一度量叫做分辨力（discrimination）。完美的测试分辨力是1，对两个结果的无用测试的分辨力是0.5，因为对角线下的区域代表没有合适模型的区域。

ROC曲线在市场营销方面的应用少于一些其他的领域。一个原因是错误肯定率很高，而错误否定率很低，即使分界得分有大的改变，曲线的形状也不会有多大改变。

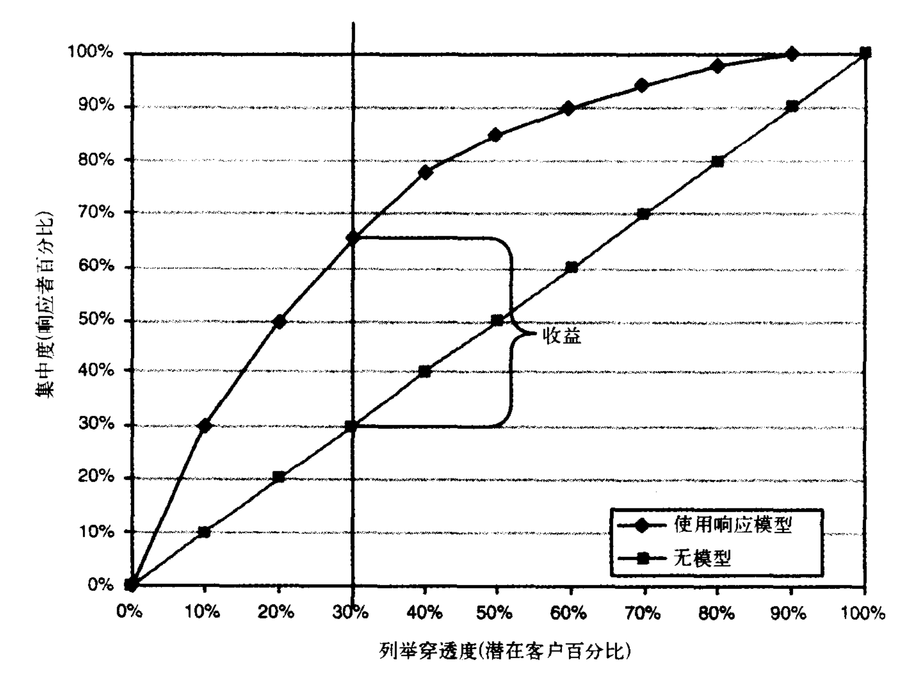


图 4-2　累积增益或集中度图表显示了使用模型的收益

上方曲线给出了集中度（响应者百分比），即越来越多的潜在客户吸引到活动中后占所有响应者的百分比。直的对角线是用于对照，它代表了未使用模型时的情况，因此集中度并不与穿透度成正比。随机抽取30%的潜在客户邮寄可能发现有30%的响应者；利用模型提供的结果，向列在前面的30%的潜在客户邮寄可能有65%的响应者。集中度与穿透度的比就是提升度，两条线之间的距离是*收益*。提升度已经在前一章讨论过，收益见“ROC曲线”部分。

这里绘出的模型在30%处有2.17的提升度，这意味着，与在100万潜在客户中随机抽取30%进行邮寄所收到的响应相比较，使用这个模型后，同样花费300 000美元，SAC将获得两倍的响应者。

4.3.3 优化营销活动收益

对于一项营销活动，能使响应率加倍无疑是一个令人期待的结果。但它实际价值到底有多大？这项活动真的有利可图吗？尽管提升度是一个有用的模型比较方法，但并不能回答上述这些重要的实际问题。要确定收益，还需要更多的信息。特别是，计算收益（profitability）不仅需要考虑收入信息，也需要计入成本信息。下面让我们对SAC例子添加一些更详细资料。

简单假设公司（Simplifying Assumptions Corporation，SAC）以单一价格销售一种产品，产品价格为100美元，生产、仓储、分发产品的总成本是55美元，另外，前面已经提到，要接触一个潜在客户需要花费1美元。现在我们有足够的信息来计算一个响应所产生的价值

了。每个响应的总值是100美元，响应的净收益要扣除与该响应相关的成本（货物成本55美元，联系成本1美元），每个响应完成的净收入是44美元。这一信息概括在表4-3中。

表4-3 简单假设公司的损益表

邮寄	响应	
	是	否
是	$44	$ -1
否	$0	$0

收益

如图4-2绘出的集中度图表（concentration chart），通常按提升度进行讨论。提升度测量集中度与穿透度的关系，把它用于比较两个模型在潜在客户列表给定深度下的性能，当然是一个有用方法。然而，它未能捕捉图中从直觉上看起来很重要的另一问题，那就是：这两条线分开有多远？在哪个穿透度处它们离得最远？

我们的同事，统计学家Will Potts，把集中度与穿透度之间的差称为收益。按他的术语，这个差值最大的点就是最大收益点。注意最大收益点与最高提升度的点并不对应。最高提升度总是在集中度图表的左边，在那个区域，集中度最大并且曲线斜率最大。

最大收益点的位置更值得关注。为了解释它的一些有用特性，本部分引用了本书正文中没有解释的一些资料（如ROC曲线和KS测试）。每一项目符号处是关于集中度曲线上最大收益点的一个命题，随后附有该命题的非正式解释。

◆ 最大收益与在每类的概率的累积分布函数之间的最大距离成比例。

意思是，在穿透度上切割潜在客户列表于最大收益处的模型分值也是Kolmogorov-Smirnov（KS）统计取最大值之处。很多统计员经常使用KS测试，特别是在金融服务行业。它是为测试两个分布是否不同而开发出来的。在最大收益点处把列表分开，产生了一个“好的列表”和一个“差的列表”，从而把“好的”和“差的”响应者的分布极大地分离，也与总人口分布极大分离。这样，“好的列表”中有最大响应者比例，而“差的列表”中响应者比例最小。

◆ 集中度曲线上的最大收益点对应于ROC曲线与无模型直线最大垂直距离处。

ROC曲线类似于常见的集中度或累积增益图，因此它们之间的这种关联并不让人感到意外。如“ROC曲线”部分揭示的那样，ROC曲线显示了两种类型的误分类差错之间的折衷。在累积增益图上的最大收益点与ROC曲线上类间距最大的点相对应。

◆ 最大收益点与灵敏度和特异性的非加权平均值最大化的决策规则一致。

如医学界所用的那样，灵敏度是在检验中得到阳性结果的人们中真阳性的比例。换句话说，就是真阳性除以真阳性与假阳性之和。灵敏度表示“一项诊断基于该检验是正确的”的可能性。特异性是在检验中得到阴性结果的人们中真阴性的比例。好的检验应该既是灵敏的又是特异的。最大收益点是能够使这两项测量值的平均值最大的那个临界点。在第8章中，这些概念改名为复检比和精度，这是在信息检索中使用的术语。复检比计算通过Web搜索或其他文本查询返回关于正确主题的文章数目；精度计算返回的文章中正确主题所占的百分比。

◆ 假定误分类成本与目标类的普及度成反比，最大收益点就与把预期损失减到最小的决策规则相对应。

评价分类规则的方法之一是对每一类型的误分类指定成本，并与基于该成本的规则相比较。无论对于响应者、缺席者、骗子，还是患有特殊疾病的人，稀少的案例通常都是最值得关注的，所以错过他们中的一个会比误分类一个普通案例成本更昂贵。按照这一假设，最大收益法会选出好的分类规则。

这个表格表明：如果一个潜在客户被联系并做出响应，该公司赚 44 美元；如果一个潜在客户被联系但没有做出响应，则公司损失 1 美元。在这个简单的例子中，选择“不和一个潜在客户联系”既没有成本也没有收益。更复杂的分析可能需要考虑这样一些事实，即不联系一个可能做出响应的潜在客户是有机会成本的，甚至作为联系的结果，通过增强商标知名度，一个未响应者可能成为好的潜在客户，并且这样的响应者可能会比单次购买客户具有更高的客户生存价值。除去这些复杂性之外，这个简单损益表可以将活动响应转变为收益图表。如果忽略活动的间接固定成本，那么即使只有 1 个潜在客户响应而另外 44 个不响应，该活动收支就平衡。如果响应率超过这个比率，该活动就有利可图。

警告：如果失败的联系成本设得过低，损益表就会建议与每个人联系。由于其他原因这可能并不是个好主意，它可能导致潜在客户被不适当的促销措施所充斥。

模型如何影响收益

如图 4-2 所示的模型中表明的提升度和收益状况会对活动的收益有何影响？答案依赖于活动的启动成本、人群中响应者的普及底线和所联系人群的穿透度边界。回想 SAC 的预算是 300 000 美元，假设人群中响应者的普及底线是 1%。该预算足以联系 300 000 个潜在客户，或者潜在客户群中的 30%。在 30% 的深度上，该模型提供了大约为 2 的提升度，因此，与没有使用模型时所能拥有的响应者相比，SAC 可以预期两倍的响应者。在这种情况下，两倍的意思是 2% 而不是 1%，产生 6 000（2% * 300 000）个响应者，他们每人的净收益值是 44 美元。在这种假设下，SAC 从响应者那里获得总收益 600 000 美元，净收益 264 000 美元。而 98% 的潜在客户或者说 294 000 人没有响应，他们每人花了 1 美元，因此 SAC 在这个活动中损失 30 000 美元。

表 4-4 显示了用来产生图 4-2 集中度图表的数据。它表明该活动可以通过花费更少的钱、联系更少的潜在客户而得到更好的响应率来赚取利润。只给 10 000 个潜在客户发邮件，即潜在客户列表的前 10%，所获提升度为 3。它将 1% 的底线响应率扭转到 3% 的响应率。在此情形中，3 000 人响应产生 132 000 美元的收入；有 97 000 人响应失败，他们每人花费 1 美元，最终总利润是 35 000 美元。更值得一提的是，SAC 在市场营销预算中剩余的 200 000 美元可以用于进行另一个营销活动，或者改善这一活动中的促销物品，因此增加的响应可能会更多。

表 4-4　以 10% 计的提升度和累积增益

穿透度	增益	累积增益	提升度
0%	0%	0%	0
10%	30%	30%	3.000
20%	20%	50%	2.500
30%	15%	65%	2.167
40%	13%	78%	1.950

（续）

穿透度	增益	累积增益	提升度
50%	7%	85%	1.700
60%	5%	90%	1.500
70%	4%	94%	1.343
80%	4%	96%	1.225
90%	2%	100%	1.111
100%	0%	100%	1.000

小型的、目标明确的活动比大型的、昂贵的活动可能获益更多。列表变短时提升度会增加。我们能因此而得出“小型的活动总会更好”这样的结论吗？回答是否定的，因为当响应者的数量减少时，绝对收入也随之减少。作为一个极端的例子，假定在底线响应率为1%时该模型能够通过发现一个100%响应率的组从而产生100的提升度。这听起来好极了，但是如果这个组只有10个人，他们仍然只值440美元。而且现实的例子应该包括预先固定成本。图4-3显示出进行如下假设时会发生什么：该活动除每联系一个人花费1美元外，还有20 000美元固定成本，每一响应收益44美元，底线响应率为1%。该活动只有在10%左右小范围的穿透度情况下是盈利的。

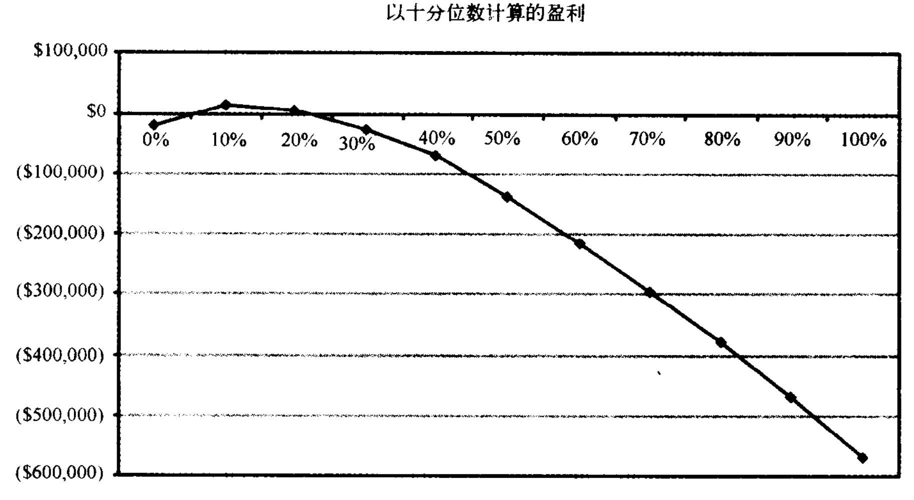


图4-3 活动收益随穿透度的变化关系

使用该模型来优化活动的收益似乎比只是用它挑选谁将包含在预定大小的邮寄或电话名单中更具吸引力，但这一方法并不是没有缺陷的。首先，最终结果依赖于活动成本、响应率和每个响应者收益，在进行该活动前其中任何一项都是未知数。虽然这些在示例中是已知的，但在现实生活中，只能大概估计。其中任何一项的很小变化将导致以上示例中的活动完全无利可图，或者要通过范围大得多的十分位数变化才能使它盈利。

图4-4显示了假设成本、响应率和收益偏离20%时该活动将会是什么结果。在最坏的情形下，能够得到的最好结果是损失20 000美元；在最好的情形下，该活动在40%穿透度处

取得最大收益 161 696 美元。成本估计趋于精确是由于事先可以确定邮资率、印刷费和其他要素，而响应率和收益估计通常是猜测。由于这些因素，虽然收益优化活动听起来很吸引人，如果预先没有进行实际试验活动，在现实中未必能实现。需要提前做的活动收益建模主要是一个基于多种假定来决定可能收益范围的假设分析。尽管预先优化营销活动不是特别有用，但在活动实施以后，用它去测试活动结果会很有用。然而要有效地做到这些，活动需要包含响应得分覆盖完整范围的客户，甚至是响应较低的十分位数的客户。

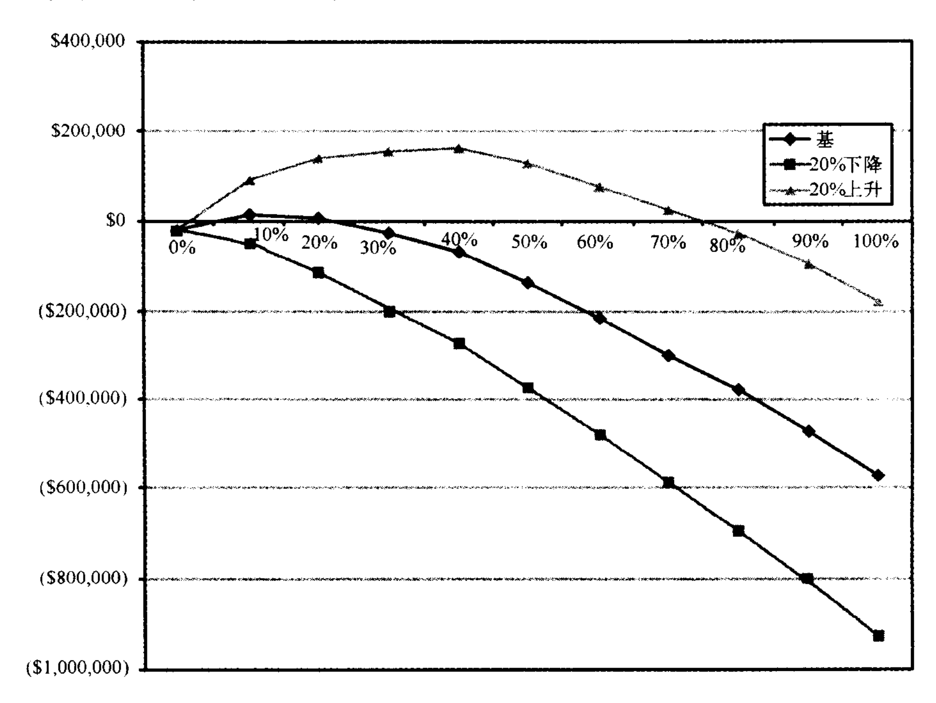


图 4-4　响应率、成本和每个响应者收益 20% 的变化对活动收益产生巨大影响

警告：由于活动收益依赖于如此多的只能事先估计的因素，惟一可靠的途径是使用现实市场进行测试。

4.3.4　接触那些受相关信息影响最大的人们

迄今为止，更微妙的简单假定之一是：具有高提升度的模型可用于识别那些对营销活动做出响应的人。既然这些人在收到一个优惠服务后就着手进行购买的比率比其他人高，该假定似乎已经被证明。然而还有另外一种可能：该模型只能简单地识别谁可能购买该产品，无论有或者没有该优惠。

这不是纯理论方面的讨论。举例来说，一家大银行进行了一项直接邮寄广告活动来鼓励客户开设投资账户，他们的分析部门开发了一个邮寄响应模型用于测试该活动，他们使用了三个群组：

- *对照群组*：随机选择的接收该邮寄的群组；
- *测试群组*：按照模型响应得分选出的应该发送邮寄的群组；
- *放弃群组*：按照模型得分选出的不宜发送邮寄的群组。

该模型给出的结果的确不错。那些具有高模型得分的客户确实比对照群组和低分客户响

应率要高。然而出乎意料的是，放弃群组中的客户响应与测试群组中的客户响应率相同。

究竟什么原因造成了这种结果？该模型工作正常，识别出了那些对这样的账户感兴趣的人们。然而，由于银行的每一部分工作都聚焦于使客户开设投资账户——播出广告、在分支机构张贴海报、在网上发消息、培训客户服务人员，所以直接邮寄被淹没在来自所有其他通道的噪声里，结果证明它是多余的。

提示：不论是一个模型还是基于该模型所推出的活动，要验证其是否有效，都需要同时跟踪考察处于没有包含在活动目标的放弃群组的潜在客户和选定为该活动目标的潜在客户二者的响应率与模型得分的关系。

市场营销活动的目标是改变客户行为。从这个意义上说，对一个无论如何都要购买的潜在客户施加影响，与对一个收到促销信息也不会购买的潜在客户施加影响相比，两者并没有太大差别。一个被识别为可能响应者的组也许同样是不太可能被市场营销信息影响的群组。他们被选定成为目标群组，表明他们在过去可能从你的竞争对手那里已经收到了许多类似的信息。他们或许已经拥有了该产品或者与之相近的替代品，或者会坚定地拒绝购买它。对于以前完全没有听说过该产品的人们，市场信息影响差别也许会更大；即使没有市场营销投资，拥有最高分的那几段可能应该做出响应。由此可以推导出近乎荒谬的结论：在市场营销投资中，响应模型中得分最高的几段也许不提供最大的回报。

4.3.5 差别响应分析

要走出这一困境，出路在于直接对活动的实际目标进行建模，这不应该只是简单地辐射那些马上进行购买的潜在客户，而是应该同时辐射那些因为被联络而更可能做出购买决定的潜在客户，这称为*差别响应分析*。

差别响应分析一般从设置一个目标群组和一个对照群组开始。如果对目标群组采取的措施具有预想的效果，目标群组的总响应将比对照群组高。差别响应分析的目的是找出目标群组和对照群组之间响应差别最大的那些群体区域。Quadstone 市场营销分析软件有一个模块，使用如图 4-5 所示的一个稍加改进的决策树来执行这一差别响应分析，他们称之为“提升分析”(uplift analysis)。

图解中的树基于从一个测试邮寄活动中得到的响应数据，如表 4-5 所示。该数据按照年龄和性别，列出了收到邮寄广告的目标群组和没有收到邮寄广告的对照群组对一项广告服务的接受率。

表 4-5 从测试邮寄活动中得到的响应数据

	对照群组		目标（邮寄）群组	
	青年	老年	青年	老年
女士	0.8%	0.4%	4.1%（↑3.3）	4.6%（↑4.2）
男士	2.8%	3.3%	6.2%（↑3.4）	5.2%（↑1.9）

无需使用数据挖掘即可看出：具有最高响应率的群组是收到邮寄的年轻男士们，随后是收到邮寄的老年男士。这是否意味着这项服务的促销活动应当主要针对男士呢？如果把目标定为“使不经促销就不会购买的新客户的数量最大化”，回答就是否定的。参与该活动的男士响应该服务的数量确实比女士多，但是更可能的情况是男士无论如何都会购买该服务。差

别响应树可以使我们更清楚地看到这样一点：受该活动影响最大的群组是老年女士。没有促销时该群组基本不会购买该服务（0.4%），但是通过促销，在购买量上她们增长了十倍。

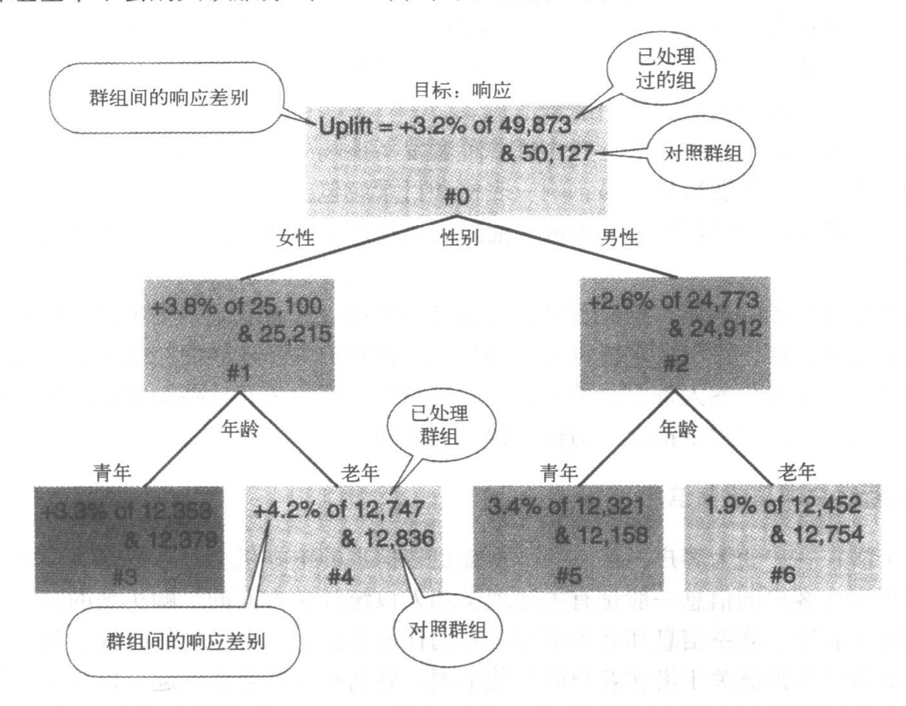


图 4-5　Quadstone 的差别响应树可以使目标群组和对照群组之间响应差别最大化

4.4　使用当前客户来了解潜在客户

发现好的潜在客户的一个好办法是察看目前最佳客户来自哪里。这意味着要有一些方法来决定谁是目前的最佳客户，它也意味着要保存这个当前客户是如何得到的以及得到时他们的特征的相关记录。

当然，依赖于当前客户来了解到哪里寻找潜在客户的不足之处在于，当前客户只反映了过去的市场营销决策。研究当前客户，将不会使你在从没尝试过的任何其他地方寻找新的潜在客户。尽管如此，当前客户的特征是一个评估已存在的获取渠道的好途径。针对发掘潜在客户这一目的，知道当前客户过去是潜在客户时的特征是重要的。最好做到以下几点：

- 在他们成为客户前就开始跟踪客户；
- 收集新客户被获取时的信息；
- 为获取时间数据和未来收益结果之间的关系建立模型。

下面几小节将提供关于这些内容的详尽讲解。

4.4.1　在他们成为客户前就开始跟踪客户

在潜在客户成为客户前就开始记录有关他们的信息是个好主意。通过网络站点能够实现这一点：每当第一次看到一个访客就发送一个 cookie 和启动一个匿名简档以记住访客做了什么，当该访客回来（使用同一计算机上的同一浏览器）的时候，该 cookie 被识别，简档被更新。当该访客最终变成客户或者是成为注册用户时，将导致这种转变的那些活动加入该客户

记录。

在离线状态下（offline world）追踪响应和响应者也是好习惯。要记录信息的第一个关键部分就是该潜在客户响应这一事实。描述谁做出了响应、谁没有做出响应的数据是未来响应模型的必要基础。只要有可能，响应数据还应尽量包含刺激该响应的营销活动、通过什么渠道做出响应以及响应是何时发出的等信息。

在诸多市场营销信息中，决定究竟哪一种激发了该响应可能是困难的，也许有时候根本就不可能。为了使工作变得更容易，响应表格和目录应包括识别代码，通过网络站点点击可以获取指向的链接。即使广告活动也应该能通过所使用的不同电话号码、邮政信箱或网站地址区分开来。

根据产品或服务的本质，响应者可能还需要为申请表或者登记表提供附加信息。如果该服务涉及信用的扩展，可能要求提供信用局信息。在客户关系建立之初收集的信息，既可能是一无所知，也可能是签人寿保险单时提供的完整医学体检表，这些初始信息各不相同。多数公司收集信息的详尽程度都介于前述两种可能之间。

4.4.2 从新客户那里收集信息

当一个潜在客户变为客户，就有一个搜集更多信息的绝好机会。在潜在客户转变为客户之前，关于潜在客户的信息一般是有关地理学和人口统计学方面的。购买来的邮寄名单不太可能提供除了名字、联系信息和名单来源之外的任何信息。当有地址的时候，就有可能通过他们邻居的特征来推测关于潜在客户的其他事项。姓名和地址合在一起可以用于从市场营销数据提供者那里购买关于潜在客户的家庭层次的信息。对于使用通常的如“年轻母亲”或“城市青少年”等分类词从大范围上锁定目标，这类数据是很有用的，但要形成个性化客户关系的基础，这种信息还不够详细。

关于最初购买日期、最初获取渠道、所响应的服务、最初的产品、最初的信用分值、响应时间和地理位置等信息收集，是将来数据挖掘会用到的最有用的内容。我们发现，对许多收益结果，如预期的关系持久度、呆账和额外购买等，这些内容具有很好的预言性。这些初始值应被保存为原始形式，不应随着客户关系的发展而被改写为新值。

4.4.3 获取时间变量可预测未来结果

通过记录在获取客户时已知的任何事项并对客户长期跟踪，利用数据挖掘技术，商家能够把获取时间变量与未来结果（如客户持久度、客户价值、隐含风险等）联系起来。然后这些信息就能够用于指导市场营销，把营销重点转向那些使该产品能够收到最大收益的渠道和信息。例如，第12章描述的生存分析技术能够用于确定每一渠道的平均客户生存期，由此通常能够发现一些渠道的客户持久时间两倍于其他渠道。假定能够大致估计一位客户的每月价值，这将转化成一个典型的渠道A客户比一个典型的渠道B客户在多大程度上更有价值的实际收益图——一个与每个响应成本测算方法同样有用的、经常用于评估渠道状况的图。

4.5 客户关系管理数据挖掘

客户关系管理自然地把重心集中在已建立的客户。已建立的客户是需要挖掘的数据的最丰富的来源。最有意义的是，已建立的客户产生的数据反映了他们实际的个性化行为：客户

是否准时支付账单？用支票还是用信用卡支付账单？最后一次购买是什么时候？购买的是什么产品？花费多少钱？该客户给客户服务中心打过多少电话？我们给该客户打过多少电话？客户最常使用什么运送方式？该客户已回头购买了多少次？这种客户行为数据能用来评估客户的潜在价值，估计他们将要结束该关系的风险，估计他们停止支付账单的风险，以及用于预测他们的未来需求。

4.5.1 按客户需求策划营销活动

用于优化针对潜在客户的邮寄预算的同一个响应模型得分对现有客户用处更多，它能被用于合理搭配由公司主导的针对现有客户的市场信息。在获取到客户以后，市场营销并没有停止，还会有交叉销售（cross-sell）活动、提升销售（up-sell）活动、使用激励活动（usage stimulation campaign）、忠诚度计划（loyalty program）等。这些营销活动可以认为是为留住客户而设计的。

若把每个营销活动孤立起来考虑，并且所有的客户在每一个活动中都设定响应分值，典型的情况是：对于其中的许多营销活动都会得高分的人处于某个相似的客户组。在该模型得分中反映出的事实是，一些客户比其他人响应更积极。这一方式会导致差的客户关系管理：高分值组被信息狂轰滥炸变得恼火而且反应迟钝，而与此同时，其他客户从未收到公司的来信，因而未受到应有的鼓励去扩展这种关系。

一种替代方案是给每一位客户寄送有限数量的信息，使用其得分来决定对每个人而言哪些信息是最恰当的。即使是对所有服务得分较低的客户也可能在响应某些服务方面得分比他人高。在 *Mastering Data Mining*（Wiley，1999）一书中，我们描述了如何用这一系统使一个金融网站更加个性化：即基于每一位客户的银行行为，对其最可能感兴趣的产品和服务高亮显示。

4.5.2 划分客户群体

划分客户群体是对已建立关系的客户进行数据挖掘的常见应用。划分群体的目的是对每一特征客户群调整产品、服务和市场推广信息。客户群体划分传统上基于市场调查和人口统计信息，比如，可能会有“年轻单身群体”或“忠诚客户群体”之分。基于市场调查划分群体的问题是，很难知道如何将这些特征应用于那些没有包括在调查中的客户。基于人口统计学划分客户群体的问题是，不是所有“年轻单身”或“居无定所者”实际上都具有所在群体的爱好和产品倾向性。数据挖掘就是通过识别行为群体进行的。

1. 发现行为群体

发现行为群体的一种方法是使用第 11 章描述的非定向聚类技术。这一方法可以产生相似客户的聚类，但要了解这些聚类与该商务的关系可能是困难的。在第 2 章中，有一个发现小商业用户群体的例子，某个银行成功使用自动聚类检测来识别好的限额家庭抵押贷款潜在客户。然而这只是已发现的 14 个簇中的一个，而其余的那些并没有明显的市场营销用途。

更典型的是，商家喜欢这样的划分群体方式：它可以把每位客户归到一个容易描述的群体中。这样一些群体经常是为续约或高消费水平等市场营销目标而建立。对于这种划分群体的方式，第 6 章描述的决策树技术是很理想的方法。

另一种常见的情形是，当预先存在基于客户行为的群体定义时，数据挖掘要解决的是在

数据中识别与群体相符的客户模式。一个好的范例是把信用卡客户分组为“频繁余额转存者”和“大额转账者”群体。

对于“发现符合预定义客户群体的模式”这一任务，数据挖掘的一个非常值得研究的应用实例是美国电话电报公司（AT&T）长途电话局用来判定电话可能用于商业用途的系统。

AT&T 把全美拥有电话并且尚非本公司客户的每一个人视为一个潜在客户。出于市场营销的目的，他们长期以来维护一个电话号码列表，称作全局列表（Universe List）。这是一个尽可能完整的美国电话号码列表，其中不仅有 AT&T 客户，还有非 AT&T 客户，每个客户被标记为商业用户或住宅用户。获取非 AT&T 客户的最初方法是从当地市话公司购买电话号码簿，搜索没有在 AT&T 客户列表上出现的号码。这样做不仅费用很高而且不可靠，随着提供号码簿的公司与 AT&T 竞争越来越直接，这种缺点变得越来越严重。

判定一个号码是住宅电话还是商用电话的原始方法是打电话询问。1995 年，贝尔实验室（那时是 AT&T 的一部分）的研究人员 Corina Cortes 和 Daryl Pregibon 提出了一个更好的办法。像其他电话公司一样，AT&T 对经过其网络的每一个电话收集通话详细数据（他们被合法授权将这一信息保存一定时间）。这些电话中有许多是由非 AT&T 客户打出或接收的。当他们拨打 AT&T 的 800 号码和从 AT&T 客户那里接听电话时，非 AT&T 客户的电话号码会出现在详细通话数据中。用已知商务活动产生的数据建立商业行为统计模型，然后把它用于对这些记录进行分析并给出其商用性分值。AT&T 称这一分值为“bizocity”，可用于判断应当给某些潜在客户推出哪种类型的服务。

每个电话号码每天都被评分。AT&T 的交换机每天处理几亿个电话呼叫，包括大约 6 500万个不同电话号码。在一个月中，他们可以看到超过 3 亿个不同的电话号码，每一个号码都被给出一个小小的档案，包括最后见到该号码以来的天数、日均使用分钟数、该号码在网络上出现的平均时间，还有 bizocity 得分。

bizocity 得分通过考虑该号码打出或接听电话的时长、一天中通话高峰时间和该号码打向已知商业电话的呼叫比率的回归模型产生，每天的新数据都会调整该得分。实际上，该得分是随时间而改变的一个加权平均值，数据越近所占权重越大。

bizocity 得分能够结合其他信息以便寻址特定的商业群体。一个过去特别感兴趣的群体是家庭商务，就连当地开通该号码的市话公司经常都不把这些看成是商务。登记为住宅地址或者被市话公司标记为住宅电话的那些具有高 bizocity 得分的电话号码，对于针对在家工作的服务是一个好的潜在客户群体。

2. 将市场调查群体与行为数据紧密联系起来

传统的基于调查的市场研究（survey-based market research）面临的很大挑战之一是，对于少数客户提供大量信息。然而，要有效地利用市场调查结果，经常需要弄明白所有客户的特征。也就是说，市场研究可以发现有趣的客户群体，然后需要将已有数据映射到现有客户群上面。行为数据对解决这种问题格外有用；这样的行为数据通常可以由转账和账单记录汇总而得到。市场调查的一个必要条件是，首先需要识别客户，以便市场调查参与者的行为是可知的。

本书中讨论的绝大多数定向数据挖掘技术都能够用于建立分类模型，然后基于现有资料将人们分派到某个群体中。为达到这个目的，需要有一个已经分类的客户训练集。这种工作的成效如何，主要依赖于客户行为对客户群体的实际支持程度。

4.5.3 减少信用风险

学会避免坏的客户（并且注意到好的客户大约要变坏的时间）与留住好的客户同样重要。大多数交易容易受消费者信用（credit risk）风险影响的公司，把进行客户信用筛选作为获取过程的组成部分，但即使在客户被获取以后，风险模型的使用也没有停止。

1. 预测谁将拖欠

对于任何客户余账服务，评估现有客户的信用风险都是一个重要问题。总会出现一些客户接受服务后未能付款这种可能性。无偿还债务是一个明显的例子；报纸订阅、电话服务、煤气和电、有线电视服务就是那些通常只有在使用之后才付款的服务例子。

当然，足够长时间都没有付款的客户最终会被终止，到那时他们或许已经欠下大量的钱但必须被一笔勾销。使用一个预言性模型的早期预警机制，公司能够设法保护自己。这些措施可以包括限制服务的使用，或者减少付款延迟与中断服务之间的时间长度。

未付款服务的终止有时称作强制流失（involuntary churn），可以用多种方式建模。在一段固定的时间范围内，强制流失经常被看做一个二元结局，像逻辑回归和决策树技术就适合于这种情况。在第 12 章中，这个问题也可看做是一个生存分析问题，实际上是将问题从“该客户下个月会不付款吗?”转换为“半数客户沦为强制流失的时间还有多久?”

自发流失（voluntary churn）和强制流失之间的一个明显区别是，在账单延迟的不同阶段，强制流失常常涉及复杂的交易过程。随着时间的推移，公司可能会收紧指导该过程的规则以控制欠款数量。当在相近的条件中寻找精确的数字时，最佳方法可能是对经营过程的每一步都建立模型。

2. 改进回收资金机制

一旦客户停止付费，数据挖掘就能在回收资金方面起帮助作用。模型用于预测能够收回费用的数量，并在某些情况下帮助选择回收策略。回收资金在根本上也可看做是某种类型的销售。公司尽力说服拖欠债务的客户支付本公司账单而不是一些其他的账单。像任何销售活动一样，一些潜在的付款者更愿意接受某种类型的信息，而有一些则更愿接受另一类信息。

4.5.4 决定客户价值

客户价值计算是相当复杂的，尽管数据挖掘会有所帮助，但客户价值计算在很大程度上是一件使财务明晰恰当的事情。客户价值似乎可以简单表述为，源于该客户的总收入减去维持该客户的总成本。但是收入中的多少该归因于一位客户？这是他（或她）迄今为止的全部花费吗？他（或她）这个月花了多少钱？我们期望他（或她）下一年花费多少？一些间接收入如广告收入和证券租赁等，应该如何分配到客户身上？

值得质疑的成本问题就更多，包括按特定方法可被分摊到客户身上的各种成本。即使忽略被分摊的成本，只看直接成本，事情可能仍然令人相当迷惑。客户并不能控制成本，那么，由于成本超支而指责客户公平吗？两个网络客户订购完全相同的商品，公司都承诺免费送货，住的地方离货场远的那个客户运输成本可能更高，但是她果真是一个价值更小的客户吗？如果另一个订单要运自不同的场所又会如何呢？移动通信服务提供者也面临类似的问题，现在大多数广告宣称全国统一费率。当他们不拥有整个网络时，这些运营商的成本就不是统一的。一些呼叫在公司网内转接，另一些可能通过竞争对手的网络转接，需要收取较高

费率。该公司可能通过试图劝阻客户不要打往某些地区来增加客户价值吗?

当所有这些问题梳理出来，并且公司对既往客户价值的定义已经协商一致，数据挖掘就可以为评估潜在客户的价值而开始工作。这归结为评估单位时间内一位客户会带来的收入以及评估该客户的剩余生存期。其中第二个问题是第 12 章中要讨论的主要内容。

4.5.5 交叉销售、提升销售和销售推荐

对于现有的客户，客户关系管理的主要着眼点是通过交叉销售（cross-selling）、提升销售（up-selling）以增加客户收益。数据挖掘用于计算应该给客户提供什么、给哪些客户提供和在什么时候提供。

1. 发现优惠的恰当时间

Charles Schwab 投资公司发现，即使在储蓄账户和投资账户有相当多的隐藏资金，客户通常仅使用几千美元在投资公司开设账户。Schwab 当然愿意吸引那样的一些资金余额。通过分析历史数据，他们发现那些把大量资金余额转移到投资账户的客户通常是在客户开设账户后的最初几个月，而几个月以后，试图使客户转入大量资金余额的努力很少会得到回报，似乎最佳窗口已经被关闭。从分析得到的这些结果，Schwab 改变它的营销策略，从花该客户整个生存期内发送恒定的诱导信息流，转变为在最初的几个月集中发送。

一家同时有每日订户和周日订户的主流报纸也注意到类似的模式。周日订户升级为每日加周日订户的现象，通常出现在客户关系建立的初期。长年累月只陶醉于周日报纸的客户根本不可能改变他（或她）的习惯。

2. 销售推荐

交叉销售的一个方法是利用第 9 章的主题——关联规则。关联规则用于发现通常可以一起出售或者倾向于被同一个人反复购买的产品簇。已经购买了簇中的一些、但并没包括全部产品的客户，对于簇中那些尚未购买的产品来说是好的潜在客户。对于零售店商品，可发现许多这样的簇，应用交叉销售方法会很有作用，但在金融服务这样的领域应用该方法则会收效甚微，因为这一领域产品相对较少，并且许多客户有相似的组合购买，这种组合购买经常是因为产品打包和以前营销工作而形成的。

4.6 保持和流失

客户流失对任何公司都是一个重要问题，对于已经远离初始指数增长阶段（initial period of exponential growth）的成熟行业尤为重要。毫无疑问，流失（或者更乐观地讲，保持）应该是数据挖掘的主要应用方面。我们使用的术语“流失”通常用于电话行业，指的是各种类型客户的自发减少或强制减少；流失（churn）是一个有用的词语，因为它既可以用作名词，也可以用作动词。

4.6.1 识别流失

给流失建模首先要面对的挑战之一是，确定什么是流失以及在出现时识别它。这在有些行业比较困难，一个极端的例子是匿名现金交易商务活动。当一个曾经忠实的客户放弃他经常去的咖啡馆，而转向同街区的另一家，熟记该客户情况的那个吧台服务员可能会注意到，但该事实将不会记录到任何公司数据库中。即使在按客户名字识别的情形下，要说出已经流

失的客户和只是有一阵没来的人之间的区别可能也很困难。假设有一个忠实的 Ford 客户，每 5 年买一辆新的 F150 敞蓬小货车，当他已经 6 年没再购买时，能因此断定他已经选择另一品牌了吗？

当有每月一次的结账关系时，比如信用卡，流失会比较容易发现。然而即使这样，客户流失多半也是悄无声息的，比如一位客户停止使用信用卡但实际没有销户的情况。流失在预订式的商务中最容易分辨，也许因为这种原因，流失建模在这些商务中最常用。长途电话公司、移动通信业务提供商、保险公司、有线电视公司、金融服务公司、因特网服务供应商、报纸、杂志和一些零售商都有一个共同的预订模式，在这种模式中的客户有正规的、需要明确终止的契约关系。

4.6.2 流失为什么重要

研究流失是重要的，因为失去的客户必须由新客户补上来，并且获得新客户的代价昂贵，而且在短期内新客户往往比已有客户带来的收益更少。这一点在市场已经相当饱和的成熟行业尤其如此——需要该产品或服务的人可能早已经从某处获得，因此新客户的主要来源是脱离竞争对手业务的那些人。

图 4-6 说明当市场变得饱和时，获取活动的响应率下降，获取新客户的成本上升。该图显示出直接邮寄活动获取每个新客户所花费的成本，假定邮寄成本 1 美元，还有以某种形式送出的价值 20 美元的优惠服务，例如一张优惠券或折扣利息率信贷卡等。该获取活动的响应率高时，例如 5%，吸纳一个新客户的成本是 40 美元（花费 100 美元去邮寄给 100 个人，其中的 5 个人响应，每人的响应成本为 20 美元，因此 5 个新客户共花去 200 美元）。随着响应率变低，则成本迅速增加：在响应率降至 1%的时候，每个新客户成本是 200 美元。从某些方面来看，与其花这些钱吸引新客户倒不如留住现有客户更有意义。

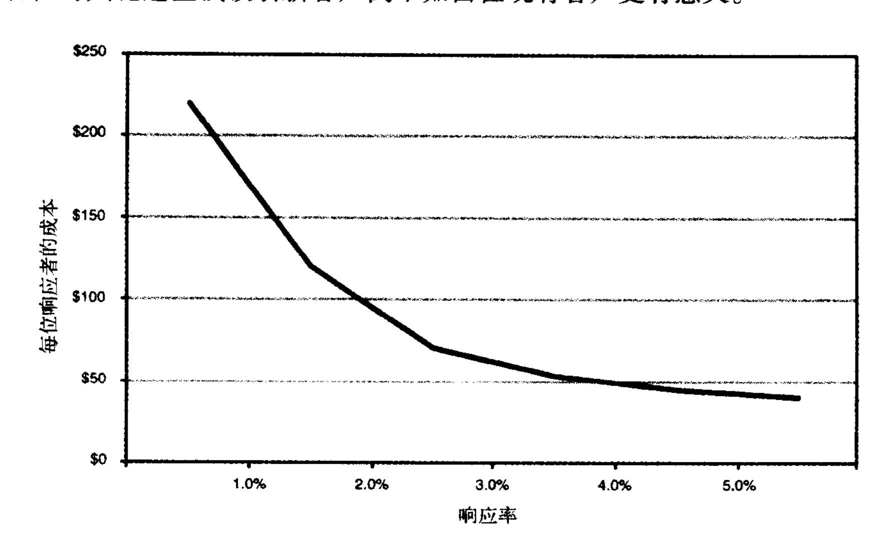


图 4-6 随着获取活动的响应率下降，获取新客户的单位成本上升

保持活动可以很有效，但也非常昂贵。移动通信公司可能会给续约的客户提供一部价格昂贵的新手机；信用卡公司可能会降低贷款利率。提供这些优惠的问题是收到该优惠的每一

位客户都将接受它。谁不想得到免费的手机或者较低的贷款利率？这意味着许多接受该优惠的人即便不给予优惠他们仍然会留下来。建立流失模型的动机是计算出谁的流失风险最大，以便对没有额外的刺激就可能离去的高价值客户提供优惠，使他们留下来。

4.6.3 不同类型的流失

对流失为什么重要的讨论实际上假定流失是自发的。客户出于自愿，决定把他们的业务挪到别的地方，这一类型的客户流失叫做*自发流失*，事实上这只是三种可能中的一种，其余两种是*强制流失*和*预期流失*。

强制流失，通常也叫*强制损耗*（forced attrition），在公司（而不是客户）终止该关系时发现——最常见的原因是由于客户未付账单。当客户不再属于一个产品的目标市场时，*预期流失*（expected churn）就会发生，比如，小孩长了牙就不再需要婴儿食品；工人退休后就不再需要养老金储蓄账户；一个家庭搬走了，就不需要再将他们订的当地报纸送到家门口，等等。

分清不同类型的流失是重要的，也很容易做到。设想在相同财务环境下的两个移动电话客户，由于某些不幸的事情，都不能再承担得起移动通信服务，因而两人都提出取消服务。一个人来到客户服务代理处办理，被记为自发流失，另一个人打电话给客服中心，等了10分钟后挂断了，然后继续使用该手机，却不付账单，第二个客户被记作强制流失。缺钱这个根本的问题对于两个客户是相同的，因此可能他们将得到相似的分值，但该模型不能预测这两个用户经历的生存期方面的差别。

把强制流失误当作自发流失的公司会有双重损失——第一次是他们花冤枉钱试图留住随后变坏的客户，第二次是新增的需要勾销的客户消费。

预测强制流失可能也是危险的。因为对不大可能付账的客户的处置令人讨厌——通信服务被暂停，滞纳金增加，催缴信发得很快。这些措施可能疏远一些也许是好的客户，并增加他们自发流失的可能性。

在许多公司，自发流失和强制流失由不同部门负责：营销部门主攻好客户的保留，财务部门则关注减少差客户的出现。从数据挖掘的观点看，不论是自发流失还是强制流失，一起处理二者会更好，因为所有的客户在不同程度上都存在这两种类型流失的风险。

4.6.4 不同类型的流失模型

流失建模有两种基本方法：第一种是把流失看做二元结局，预测哪些客户将会离去，哪些会留下来；第二种方法是设法评估客户的剩余生存期。

1. 预测谁将离去

把流失作为二元结局建模，需要选取一定的时间范围。如果问题是“明天谁将离去?”其答案几乎不会是任何一位客户。如果问题是“100年后谁将离去?”在绝大多数商务活动中，其答案几乎是每个人。二元结局流失模型通常有一个相对较短的时间范围，比如60天或90天。时间范围当然不能太短，否则将没有时间按照模型预测来采取行动。

包括逻辑回归、决策树和神经网络在内的任何常见的分类工具都能够用来建立二元结局流失模型。可以把描述一位客户的历史数据与显示这位客户在随后一段时间是否仍然活跃的标志结合起来。建模的任务是把即将离去和留下的客户区分开来。

二元流失模型给出的典型结果是一个可以按流失可能性对客户分级的分值。最常见的分值只不过是客户在该模型时间范围内将要离开的可能性。超出某一自发流失分值阈值的那些客户可以被划到保持计划中，而超出某一强制流失分值阈值的那些客户可以被放置到观察列表中。

典型地，对流失的预测既要关注该客户在获取时已知的事件（例如获取渠道和初始信用等级），还要关注在客户关系中出现的事件（例如服务问题、逾期付款和出乎意料的高账单或者低账单）。第一类流失提供的信息是如何减少获取倾向于流失的客户，以便降低未来的流失，第二类流失用于提供如何对已存在的客户减少流失风险的洞察力。

2. 预测客户将要停留多久

流失建模的第二种方法虽然具有一些吸引人的特征，却并不常用。这种方法的目标是计算出客户可能会保持多长时间，这比简单地说该客户是否将在 90 天内离开提供的信息更多。对客户剩余生存期的估计是建立客户生存期价值模型的必要条件。它可能也是客户忠实度分值的基础，该分值把忠实客户定义为在未来将长期保持的人，而不是到现在为止已经保持了很长时间的人。

对客户生存期建模的一种方法是拍下现有客户群的快照，连同这些客户最初被获取时的特征数据，通过发现那些较早获取的长期客户具有的共同特征来直接估计客户保有期。这种方法的问题是，圈进的客户时间越长，他们被获得时的初期市场环境与当今环境相比差别越大。比如，假定把 1990 年签订移动电话入网协议的某人的特征作为今天新客户将保持服务的预测器，这种做法当然不可靠。

一个更好的方法是使用从统计学中借鉴并加以改进的生存分析技术，这些技术在医疗领域被用于研究病人在医学干预后的生存率，在生产领域被用于研究部件的预期损坏时间等。

生存分析将在第 12 章中进行讲解，基本思想是计算每位客户（或者是具有相同地理学、信用等级和获取渠道等模型输入变量值的一组客户）迄今为止进展正常，但将要在明天之前离去的概率。对任意一个保有期而言，这种风险性都是非常小的，但对某些保有期会比其他一些高。客户将继续存在直到某一更远的未来日期的可能性，可以从干预风险计算出来。

4.7 小结

在包括生物工程研究和制造过程控制在内的各个领域中，本书描述的数据挖掘技术都有应用。然而，本书的目标读者是那些像作者一样，将这些技术应用于在市场营销和客户关系管理中出现的各种商业问题的人们。本书绝大部分选用的阐明某些技术的示例中，都隐含了以客户为中心的应用目标，这一点在本章中更为明显。

数据挖掘可以用于广告和定向市场营销，以识别正确的受众、选择最佳沟通渠道和挑选最适当的信息。潜在客户可以与预期受众的简档相比较并给出匹配度得分。即使不知道潜在客户个体的信息，利用美国人口普查局、加拿大统计署和许多国家的类似官方机构来源的这类数据，通过同样的方法也能为地理上的邻居给出匹配度得分。

数据挖掘在定向建模方面的一个重要应用是响应建模。响应模型给出潜在客户响应定向市场营销活动可能性的分值。这一信息能够用于改善活动的响应率，但是仅靠这一点不能判定活动的收益。评估活动收益需要依靠对未来活动的潜在响应率估计、与响应相联系的平均订购数量估计、执行活动以及活动本身的成本估计。一个更多以客户为中心的响应分值的用

途是从许多竞销活动中为每一客户选择最佳活动。这可以避免一个常见问题，即那些每次都选出同一些人、各自独立的、基于得分的营销活动。

一个模型可能有识别对某个产品或服务感兴趣的人的能力，也可能具有识别由于被某个特定营销活动或优惠吸引而进行购买的人的能力，把模型的这两种识别能力区分开来是非常重要的。差别响应分析提供了一个方法，可用于识别活动将有最好效果的市场群体。差别响应模型的目标是，在目标群组和对照群组之间，寻求把响应的差别最大化，而不是试图将响应本身最大化。

从当前客户成为客户之前的已知信息中找出目标结果的预测值，利用当前客户的信息可以识别出可能的潜在客户。这种分析对于选择获取渠道和联系策略以及筛选潜在客户列表是有价值的。公司能够通过从客户第一次做出响应，甚至在他们成为客户之前，就开始跟踪他们，并在获得客户时收集和存储附加的信息，以此增加客户数据的价值。

一旦获得客户，公司的工作重点就转换为客户关系管理。现有客户的可用数据比潜在客户的可用数据更丰富，由于这些数据本质上比单纯的地理和人口统计学信息更具行为科学性，因而它具有更好的预言性。基于客户当前使用模式，数据挖掘可用于发现应当提供给他们哪些额外的产品和服务，也能对交叉销售和提升销售的最佳时机提出建议。

客户关系管理计划的目标之一是留住有价值的客户。数据挖掘能帮助识别哪些客户最有价值，以及评估与每一客户相关联的自发流失或强制流失风险。掌握了这些信息，公司能将优惠服务锁定于既有价值又具流失风险的客户，并采取相应措施避开可能违约的客户，保护自己。

从数据挖掘的观点看，流失模型的建立既可以作为二元结局预测问题，也可以通过生存分析来解决。这两种方法各有利弊：二元结局方法对于短期情况工作良好，而生存分析方法可用于对未来远景做出预报，并且提供对客户忠诚度以及客户价值的洞察。

第5章　统计学的魅力：数据挖掘常用的工具

从统计学家（或经济学家）的角度来看，数据挖掘长期以来带有某种贬义。似乎数据挖掘不是从大量的数据中发现有用的模式，而是有寻找数据以适应某种成见的嫌疑。这很像政客围绕选举的所为——搜寻数据来证明他们的政绩和成功；这当然不是我们所指的数据挖掘的含义！本章意欲在统计学家和数据挖掘者之间的隔阂上搭建桥梁。

统计学和数据挖掘这两个学科的确非常相似。统计学家和数据挖掘者通常使用许多相同的技术，现在统计软件厂商在他们的软件包中包含许多下面8章中将讲到的技术。150多年来，统计学作为一门从数学中分离出的学科，帮助科学家理解观察的意义，设计与科学方法相联系的、输出可重复的精确结果的实验。几乎在所有这段时间内，存在的问题不是数据太多，而是太少，为此科学家们不得不用在笔记本里手工收集的数据进行计算，以便更好地了解这个世界。但这些数值有时被错记，有时由于褪色或墨渍而难以辨认，给计算造成了更多困难。早期的统计学家是从事实际工作的一些人，他们发明了一些技术用于处理手头遇到的各种问题。如今统计学家仍然是有实践经验的人，他们使用现代技术，也使用历经实践证明可靠的技术。

不同寻常且值得告慰现代统计学奠基者的是，在极少量数据上发展出来的那些技术存活下来，并且仍然被证明是有效的。这些技术不仅在最初的那些应用领域，而且实际上在所有存在数据收集的领域，从农业到心理学、天文学乃至商业，都证明了它们的价值。

或许20世纪最伟大的统计学家是R. A. Fisher，他被许多人尊为现代统计学之父。在20世纪20年代，现代计算机发明之前，他发明了设计和分析科学实验的方法。在伦敦郊外农场生活的两年时间里，他收集了各种各样的农作物产量连同其潜在的解释性变量——例如雨水、阳光和施肥量等。为理解什么对作物产出有影响，他发明了新技术（例如方差分析——ANOVA），并对收集的数据进行了上百万次计算。21世纪计算机芯片毫不费力就能在一秒钟内处理许多个百万次运算，而Fisher的每一次计算都需要在手动计算机器上拉动控制杆，伴随疼痛的双手和老茧，经过日积月累一点一滴获得结果。

计算机的出现已经明显地简化了分析的一些方面，尽管它的更大作用或许是产生了大量的数据。我们的目标不再是从每一个珍贵的数据中萃取可能的少量结果信息，而是变为理清如此海量数据的意义，因为在原始格式下存在的这些数据已经远远超出我们头脑自身的理解能力。

本章的目的是介绍统计学的一些关键思想，它们已经被证明是数据挖掘的有用工具。这种介绍的定位，既不是全面的统计学介绍，也不是泛泛的阐述；相反，它是对一些有用的统计学技术和思想的介绍。这些工具用实例进行展示，而不是通过数学证明。

本章开始于怀疑态度的介绍（这或许是应用统计学最重要的一个方面），然后讨论如何透过统计学家的眼光来考察数据，沿着这一思路介绍重要的概念和术语。本章穿插讲述一些应用实例，尤其是置信区间（confidence interval）和卡方检验（chi-square test）。最后一个示例，使用卡方检验来理解地理布局和渠道，是本章所呈现思想的一个与众不同的应用。本章

最后以数据挖掘者与统计学家之间的区别的简要讨论结束——他们态度方面的差别仅仅是程度上的而不是实质上的。

5.1 Occam的剃刀

William of Occam是一个圣芳济会的修道士，1280年出生在英格兰的一个小镇——这个时间不单是在现代统计学发明之前，甚至是在文艺复兴和印刷术之前。作为一位有影响的哲学家、神学者和教授，他向人们阐述了关于事物的许多思想，包括教堂政治。身为一个修道士，他是一个恪守贫穷誓约，过着严格自律生活的禁欲主义者；他也是合理权力的热烈倡导者，否认绝对真理的存在，而且赞成一种现代哲学思想，这种思想非常不同于大部分生活在中世纪的同时代人对生活的看法。

William of Occam与数据挖掘有什么关系呢？他的名字已经与一个非常简单的思想结合起来。他自己用拉丁语（有学问的语言，甚至那时英国人也这样认为）解释了它，“Entia non sunt multiplicanda sine necessitate”。用我们比较熟悉的话，我们会说“越简单越好”，或者通俗一点，“使它保持简单，傻瓜。”任何解释应该做到：努力使原因的数目变成一个尽可能的最小量，这种推理的思路被称为Occam的剃刀（意思是理个光头是最简单的），这也是William of Occam对数据分析的贡献。

William of Occam的故事有一个很有意思的结局：也许因为他对合理权利的追求，他也相信教堂的权力应该与国家权力分开——教堂应该仅限于宗教性的事务。这导致了他反对罗马教皇约翰二十二世干预政治，最后自己被逐出教会。他最后在1349年瘟疫爆发期间死于慕尼黑，留给后代世人的遗产就是一种有条理的和批判性的思考方式。

5.1.1 原假设

Occam的剃刀对数据挖掘和统计学是非常重要的，虽然统计学表达该思想有一点不同。*原假设*（null hypothesis）是假定在观测中的不同只归因于偶然性。举例来说，假设有一个总统选举民意测验结果，候选人A得票率45%，候选人B得票率47%。因为这一数据来自民意调查，可能有一些错误的来源，因此，数值只是每位候选人受欢迎程度的大约估计。外行可能会问，“这两个数值不同吗？”统计家对这个问题的提问会稍微有些不同，“这两个数值真正相同的概率（probability）有多大？”

虽然两个问题非常相似，但是统计学家表现了对该问题的一点态度。这种态度就是，这点差别可能根本并不重要，它是一个原假设的例子。这一例子看上去虽有2%的差别，然而这一观察到的数值可能由被响应人的特定样本来解释。取另外一个样本有可能给出相反的2%的差别，或者可能有0%的差别，这都是民意调查相当可能的结果。当然，如果倾向有20%的差别，那么它可能很少会是由抽取样本的差异造成的。如此大的差别将大大增强一位候选人比另一位做得好的可信度，并大大减少原假设成真的可能性。

提示：最简单的解释通常是最好的——即使（或尤其）在它没有证明你想证实的假说的时候。

这种怀疑态度不论对统计学家还是数据挖掘者都是非常有意义的。我们的目标是展示确实起作用的结果，并尽量减少原假设。数据挖掘者和统计学家之间的一个差别是，数据挖掘者时常面对足够大量的数据，没有必要去考虑那些归因于偶然性事件的概率计算技巧。

5.1.2 p值

原假设不仅是一种分析方法，它也能够被定量，一般常用 p 值给出原假设为真的概率。记住，当原假设为真时，表示真的没有发生什么，因为差异归因于偶然性。许多统计学家一直在致力于确定 p 值的界限。

考虑前面总统民意测验的例子。设想 p 值被计算为 60%（关于这是如何计算的更多内容将在本章稍后讨论），严格地说，这意味着有 60%的可能性是：民意测验给出的对两位候选人的支持率差别主要是由于抽取样本的偶然性引起的，而非真的是由于大众总体的支持情况造成的。在此情况下，几乎没有证据表明对两位候选人的支持是有差别的。

让我们来看 p 值改为 5% 的情形。这是一个相对来说很小的数字，它意味着，我们有 95%的信心认为候选人 B 比候选人 A 做得好。置信度，有时称为 q 值，是 p 值的反面。通常的目标是追求至少 90%的置信层次，如果达不到 95%或者更多的话（这意味着相应的 p 值分别小于 10%或者 5%）。

这些概念——原假设、p 值和置信度——是统计学的三个基本概念。下一小节将详细阐述这些概念并介绍统计分布的概念，还将特别介绍正态分布（normal distribution）的概念。

5.2 观察数据

统计是指在抽样数据上进行的测度，统计学就是对这些测度和被测度样本的研究。因而，介绍统计学的一个好的起点应该是从这些有用的测度和如何观察数据开始。

5.2.1 观察离散数值

用于数据挖掘的许多数据实际上是离散的，而不是连续的，这些离散数据以产品、渠道、区域和有关商务的描述性信息表现出来。这一节讨论观察和分析离散字段的方法。

1. 直方图

关于离散字段的最基本描述性统计是不同数值出现的次数。图 5-1 显示了一段时间内“停止”理由代码的直方图（histogram）。直方图显示出每个数值在数据中出现的频繁状况，既可以用绝对次数（204 次），也可以用百分数（14.6%）来表示。通常有太多的数值要显示在单个直方图中，例如这个案例中有超过 30 种另外的代码被分组到“OTHER”类中。

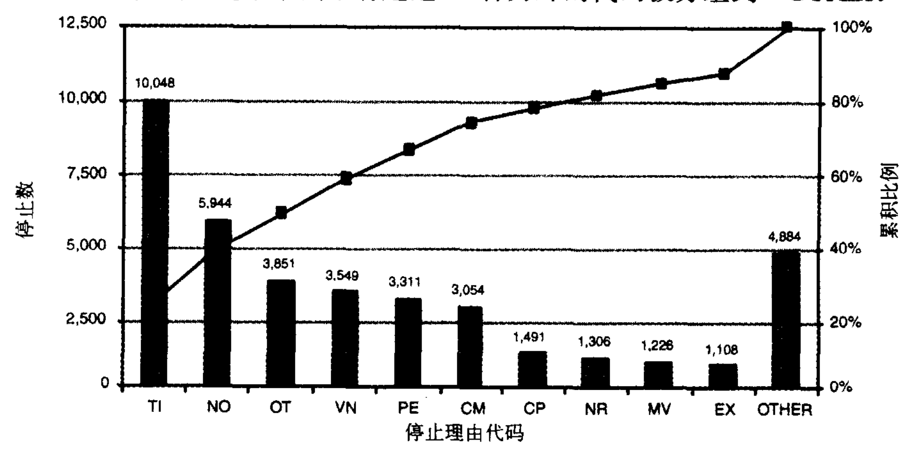


图 5-1 这个示例在同一图表中同时给出关于特定市场营销工作的停止理由的直方图（作为竖直条形图）和累积比例（作为折线）

除了每一个类的数值之外，直方图也显示了停止的累积比例，其标度列示在左侧。通过累积直方图，可能看到出现频率最高的前三个代码占停止理由的大约 50%，而前 10%的代码则几乎包含了 90%的停止理由。作为有美感的注解，表格线在敏感的点上横贯左右两边的侧标尺，使从图中读取数值更容易。

2. 时间序列

直方图很有用，使用 Excel 或任何统计软件包都能很容易做出来，然而直方图描述的只是单一时刻。数据挖掘经常关心随时间的推延会发生什么，一个关键问题是：随时间流逝，数值的频率是否恒定不变。

时间序列分析（time series analysis）需要对数据选择适当的时间帧。这不仅包括时间的单位，还包括我们从何时开始计数。几个不同的时间帧（time frame）的例子是：顾客关系开始的时刻，顾客请求停止的时刻，实际的停止日期等诸如此类的时刻。不同的字段属于不同的时间帧，例如：

- 描述顾客关系开始的字段——如初始产品、初始渠道或初始市场——应当从顾客最初开始日期着手；
- 描述顾客关系终止的字段——如最后产品、停止理由或停止渠道——应当在顾客关系终止之日或者顾客的保有期的那一点及时观察；
- 描述顾客关系期间事件的字段——如产品升级或降级、对促销的响应或滞纳付款——应当在事件发生之日、顾客的保有期的那一时刻或者在某些其他事件之后的相对时间等时刻来观察。

下一个步骤是标绘图 5-2 所示的时间序列。这个图按照停止日期有两个停止序列：一个显示的是某一特定停止类型（提价停止）随时间的变化，另一个是停止的总数。注意时间轴的单位是以天计的，尽管许多商业报表是在周报和月报水平完成的，我们还是喜欢按日观察数据，以便看到在精细水平下可能出现的重要模式，或者那些通过汇总可能变模糊的模式。在这一案例中，两条线都有清楚的上升和下降摆动模式，这是由于停止是以每周作为周期观察的。此外，浅颜色的线条描述的是价格增长相关的停止，该图清楚地表明，由于定价的改变，2 月份开始停止出现显著增长。

提示：当观察一段时间内某字段的值时，按日观察数据可以得到最细粒度水平上数据的感觉。

时间序列图包含很多信息。例如，根据数据做出一条直线可以查看和量化长期趋势，如图 5-2 所示。因为季节性的原因，这样做时要特别小心。使用非整年份的信息有可能产生某种片面性的趋势，因此使用最佳匹配线时应该包括整年数据。这个图中的趋势显示出停止的增加，但这也许无需担心，因为在这段时间内，顾客数量也在增长。这提示我们，更好的测度方式应当是停止率，而不是停止的原始数量本身。

3. 标准值

时间序列图提供了有用的信息，但是并没有解释这种随时间的变化是预期的还是意料之外的。为达到这个目的，需要利用一些统计学工具。

观察时间序列的一种方式是把所有这些数据作为一个分区，每天一点点来观察。统计学家现在要问一个怀疑性的问题："可以把每天看到的差异完全归因于偶然性吗?" 这是一个原假设，可以通过计算 p 值（数值之间的偏差能单纯用偶然性解释的概率）来回答。

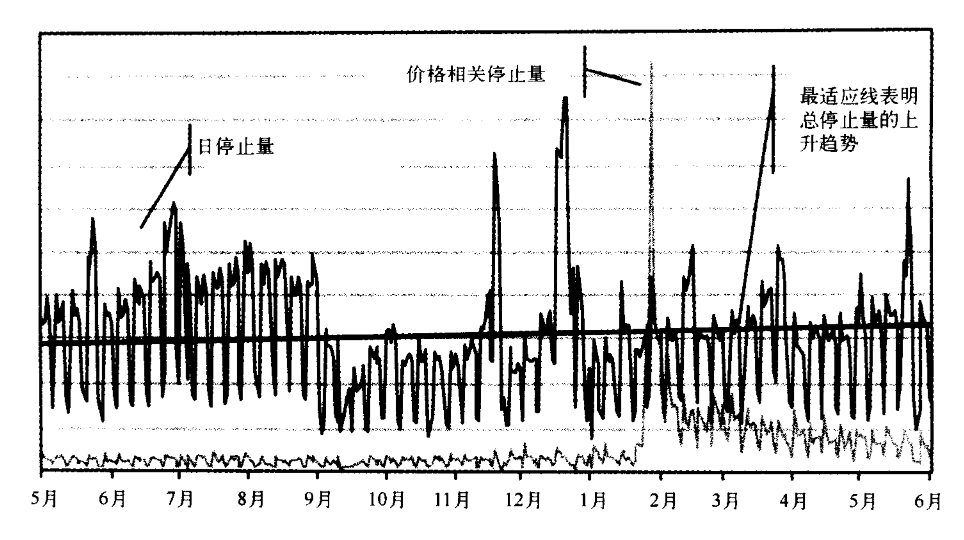


图 5-2 本图显示了用不同标度描绘的两个时间序列。深颜色的线表示总停止量，浅颜色的线表示价格相关因素停止量，它显示出 1 月末由于价格策略改变而产生的影响

统计学家对这一基本问题已经研究了一个多世纪。幸运的是，他们也找到解决这个问题的一些方法。这是一个关于样本差异的问题，每一天代表一个取自该期间所有停止的停止样本，在不同日期观察的停止偏差可能简单地归因于在随机抽样过程中的预期偏差。

在统计学中有一个基本定理，叫做"中心极限定理"，表达如下：

> 当从总体中取出的样本越多时，样本平均值（或类似统计量）的分布越接近正态分布。样本的平均值（统计学家称它为均值）肯定会越接近全部总体的平均值。

中心极限定理实际上是一个很高深的理论并且很有趣，更重要的是它很有用。就离散变量而言，比如每一天停止的客户数量，具有同样的规律。用于这一示例的统计量为每日停止计数，如图 5-2 所示。（严格说来，使用比例更好，如停止数量与客户数量的比例；在假定该段期间内客户数量恒定的前提下，这与我们所用的计数是等效的。）

正态分布由两个参数描述，均值和标准差（standard deviation）。均值是每一天的平均计数；标准差是数值趋向于均值聚集程度的度量，这将在本章后面进行更详细的解释，眼下知道可以使用一个函数（如 Excel 中的 STDEV（）或 SQL 中的 STDDEV（））进行计算就足够了。对于该时间序列，标准差是指每日计数的标准差。假定每天的数值是从整个时期的停止中随机抽取，计数的集合应当遵循正态分布；如果它们不遵循正态分布，那就是除了偶然性之外还有其他因素影响该数值。注意这并没有告诉我们到底是什么在影响该数值，仅仅利用一个最简单的解释——样本差异——是不足以解释它们的。

这就是标准化时间序列数值的目的之所在，这一过程从平均值中产生标准差的数值：

- 计算全部日期的平均值。
- 计算全部日期的标准差。
- 对每个值，减去平均值并除以标准差以得到距离平均值的标准差的数值。

标准化这些数值的目的是测试原假设。当正确时，标准化值应当遵循正态分布（均值为 0 并且标准差为 1），且显示出以下几个有用的性质。首先，标准化的值应当取以大体相等的频率出现的负值和正值。同样，当标准化后，大约 2/3（68.4%）的数值应当在 −1 和 1 之

间，略多于 95% 的数值应在 −2 和 2 之间，大于 3 或小于 −3 的数值应当非常稀少——或许在数据中根本看不到。当然，这里的“应当”是指数值遵循正态分布并且原假设有效（即全部时间相关影响均由样本差异解释）的前提下，当原假设无效时，从标准化值常可以显而易见地看出来。“术语的问题”部分谈了关于分布的更多内容——正态和非正态分布。

图 5-3 显示了图 5-2 中数据的标准化值。首先应注意到的是标准化曲线的形状与原始数据的形状非常相似，改变的只是纵轴的标度。比较两条曲线可以发现，各自的标度都发生了改变。在前一个图中，总体停止值比定价停止值大得多，因此两者使用不同的标度显示。在这一个图中，标准化的定价停止值大大高出标准化的总体停止值，尽管两者使用了同样的标度。

图 5-3 中的总体停止是十分典型的正态分布，但以下几点需要说明。在 12 月份有一个大的峰值，这可能需要解释，因为该值偏离平均值超过四倍标准差；还有，该图呈很强的以周为周期的变化趋势，也许用每周停止值取代每日停止值来重画这个图，在每周的水平上看待变化是个好主意。

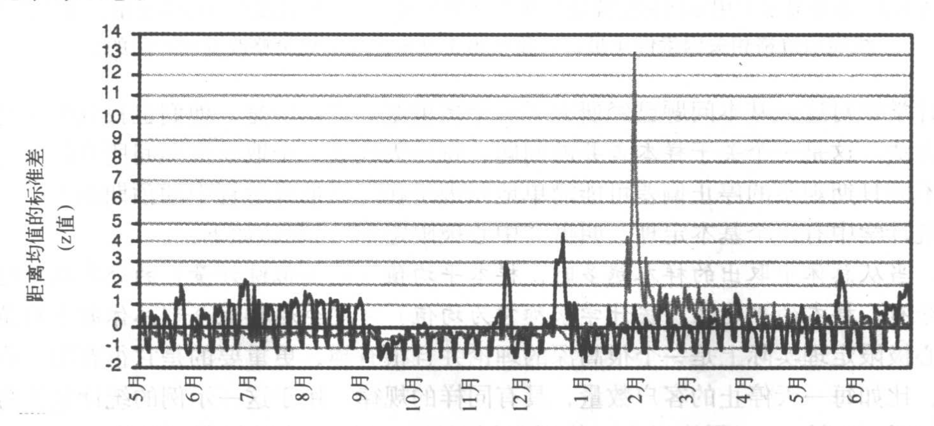


图 5-3　标准化值使我们可能在同一图表中使用相同标度对比不同的组；这个图显示总体停止数和价格增长相关的停止数

显示价格相关停止值的较浅线条明显不遵循正态分布：负值比正值多得多，峰值出现在高于 13 处，远远高出很多。

标准化值（或通常称作 z 值）是十分有用的。这一示例利用它们观察时间段数据，来看这些值是否像是从每一天随机抽取的，亦即，是否每日数值的差异能够用样本差异来解释。在当 z 值相对高或低的日期，我们怀疑有别的因素在起作用，有其他因素影响停止的出现。例如，在定价停止中，峰值的出现是因为定价的改变，在每日的 z 值中，这种影响是相当明显的。

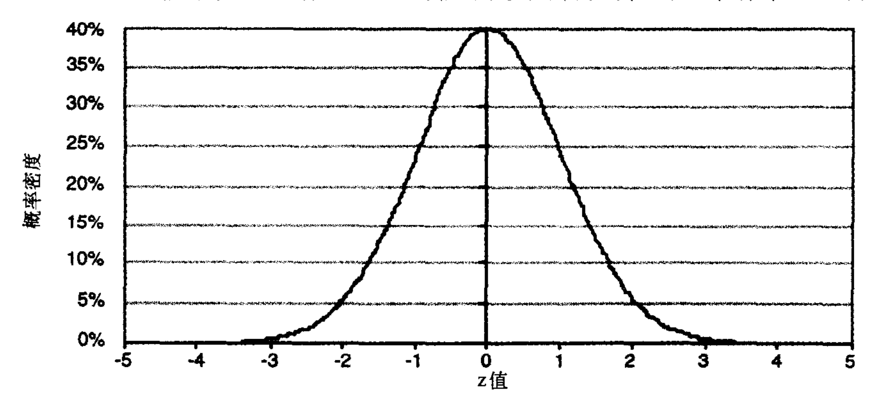
z 值也因为其他原因很有用处。例如，它是取得几个变量并将其转化为相似范围的一种方法，这对几种数据挖掘技术（例如聚类和神经网络）都很有用。z 值的其他用途在第 17 章讨论数据转换时会讲到。

术语的问题

统计学中一个很重要的观点是分布的观点。对于离散变量，分布很像直方图——它表明一个给定值以 0 到 1 之间的概率出现的频度。例如，均匀分布表示所有值是均等出现的。均匀分布的一个例子是在顾客用信用卡支付的商业活动中，用美国万国宝通卡、维萨卡和万事

达卡支付的顾客数量同样多。

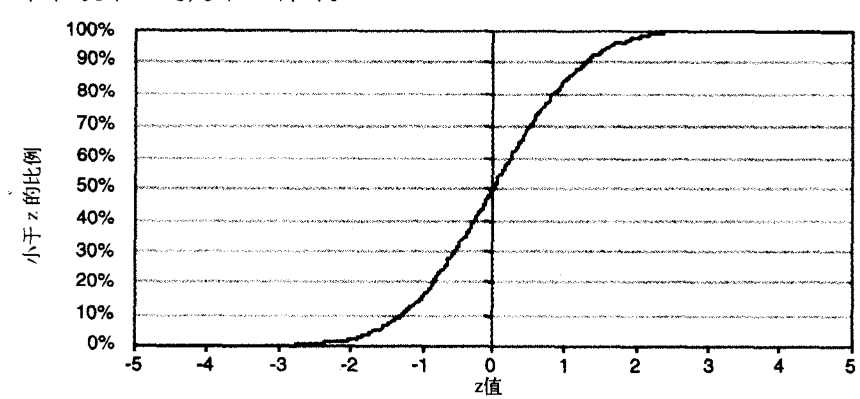
在统计学中扮演十分特别角色的正态分布是连续变量分布的例子。下图显示的就是正态（有时称为高斯或钟型）分布，其均值为0、标准差为1。读懂这一曲线的方法是观察两点之间的区域，对于遵循正态分布的值，该值落在两个值之间（例如在0和1之间）的概率是曲线下的面积。对于值0和1，其概率为34.1%；这意味着一个遵循正态分布的变量，有34.1%的时间将在大于均值的一个标准差之内具有某一个值。因为曲线是对称的，在小于均值的一个标准差内另外还有34.1%的概率，因此有68.2%的概率处于偏离均值的一个标准差之内。



正态分布的概率密度函数看上去像是常见的钟形曲线

上面给出了一个钟形曲线图，我们称之为正态分布。实际上，确切的术语应该是密度函数（或概率密度函数）。尽管这一术语衍生自高等数学概率论，但它是有意义的。密度函数给出一个变量有多“密”的程度。我们通过度量两点之间曲线下的面积来计算密度函数，而不是通过读取单独的数值本身。在正态分布的情形下，数值密集在0的周围，并且随着距离远离0而密度越来越稀疏。

下图显示了真正呈正态分布的函数。这种变化范围从0到1的形式也叫做累积分布函数。数学上，值X的分布函数被定义为变量具有小于或等于X的值的概率。因为“小于或等于”这个特性，这一函数总是从0附近开始，向上攀升，并到1附近终止。一般来说，密度函数给人们提供了关于分布情况的更直观提示。因为密度函数提供了更多信息，经常也称为分布，尽管在技术上这是不正确的。



呈正态分布的（累积）分布函数呈S形并绕Y轴呈反对称

4. 从标准化值到概率

标准化值遵循正态分布的假定使得计算某个数值偶然出现的概率成为可能。实际上，所用的方法就是计算某远离均值的事件出现的概率，即 p 值。确切数值不值得深究，因为任何给定的 z 值都有一个任意小的概率。概率被定义为在 z 值范围内正态曲线下两点之间的面积。

计算某事物远离均值可能意味着以下两种情况之一：

- 距均值多于 z 个标准差的概率。
- 比均值大 z 个标准差（或者比均值小 z 个标准差）的概率。

第一个称为双尾状分布（two-tailed distribution），第二个称为单尾状分布（one-tailed distribution）。该术语从图 5-4 中可以看得很清楚，因为分布的拖尾正在测量。对 z 值，双尾状概率总是单尾状概率的两倍，因此，双尾状 p 值比单尾状者更保守；就是说，双尾状更可能假定原假设为真。如果单尾状给出原假设的概率为 10%，那么双尾状给出的将是 20%。作为约定，为保险起见，使用双尾状概率计算更好。

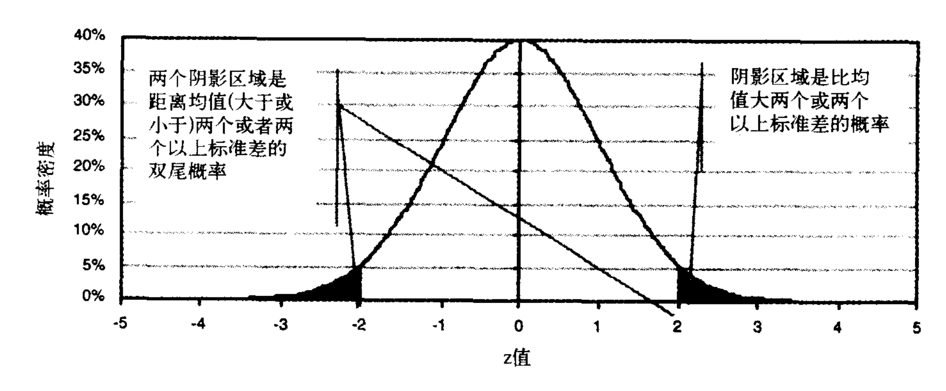


图 5-4 正态分布的尾部回答以下问题："得到一个值为 z 或大于 z 的概率是多少?"

双尾状分布的 p 值能够在 Excel 中很方便地计算出来，因为有一个叫做 NORMSDIST 的函数计算累积正态分布。使用这一函数，双尾状 p 值等于 2 * NORMSDIST（-ABS（z））。对于值 2，结果为 4.6%。这意味着有 4.6% 的概率观测到一个值超出平均值两个标准差——从平均值加或减两个标准差。或者换一种说法，有 95.4% 的置信度说明，一个值落在两个标准差的外部是由于偶然性之外的事件引起的。对于精确到 95% 的置信度，可以用 1.96 倍的界限替代 2；对于 99% 的置信度，该界限为 2.58。下列各项显示了一些通用置信层次（confidence level）下 z 值的界限：

- 90%置信度→z 值>1.64
- 95%置信度→z 值>1.96
- 99%置信度→z 值>2.58
- 99.5%置信度→z 值>2.81
- 99.9%置信度→z 值>3.29
- 99.99%置信度→z 值>3.89

置信度有如下性质：当值不可能是由于偶然性引起时，它接近于 100%；当归因于偶然性时它接近于 0。有符号（正或者负）的置信度增加了关于该值是过低还是过高的信息。当观测值低于平均值时，有符号的置信度为负值。

图 5-5 显示了图 5-2 和图 5-3 中所示数据的有符号的置信度，使用双尾状概率。有符号的置信度形状与早期的形状不同。“停止”总体各处摆动，通常保持在合理的界限以内。然而定价相关停止再一次呈现出非常独特的模式，在很长时间内是很低的，剧烈增加后又下降。有符号置信度水平界限为 100% 和 - 100%。在本图中，极值接近 100% 或 - 100%，且很难说出在 99.9% 和 99.99999% 之间的区别。要区别接近极端的数值，图 5-3 中的 z 值比有符号的置信度更好。

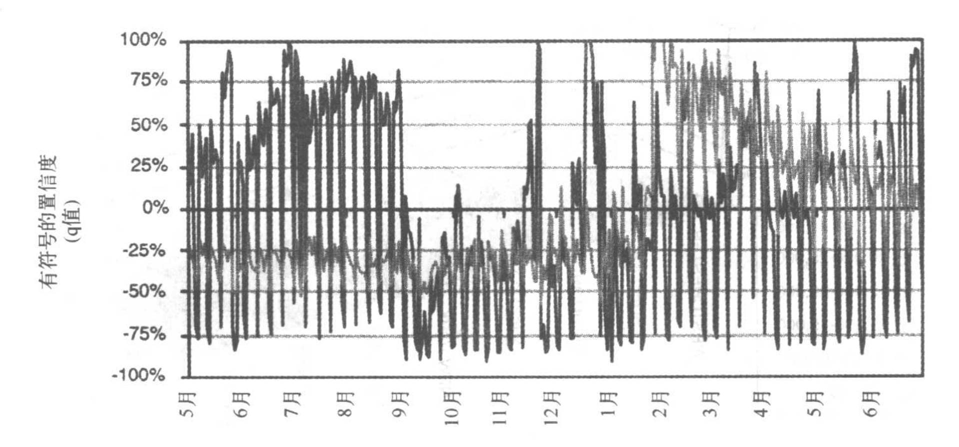


图 5-5　以图 5-2 和 5-3 相同的数据为基础，该图显示了基于平均值和标准差的被观测值的有符号置信度（q 值）。当被观测值过高时此符号为正，过低时为负

5. 交叉表

时间序列是交叉表（cross-tabulation）的一个例子——同时观察两个或更多变量的值。对于时间序列，第二变量为某事件发生的时间。

表 5-1 显示了本章稍后使用的一个例子，该交叉表显示出如下三种渠道下来自纽约州东南部郡县的新客户数量：“电话销售”、“直接邮寄”和“其他”。表中不仅显示原始计数还显示相对频率。

表 5-1　按郡县和渠道显示的交叉表

郡　县	计　数				频　率			
	电话销售	直接邮寄	其他	合计	电话销售	直接邮寄	其他	合计
BRONX	3 212	413	2 936	6 561	2.5%	0.3%	2.3%	5.1%
KINGS	9 773	1 393	11 025	22 191	7.7%	1.1%	8.6%	17.4%
NASSAU	3 135	1 573	10 367	15 075	2.5%	1.2%	8.1%	11.8%
NEW YORK	7 194	2 867	28 965	39 026	5.6%	2.2%	22.7%	30.6%
QUEENS	6 266	1 380	10 954	18 600	4.9%	1.1%	8.6%	14.6%
RICHMOND	784	277	1 772	2 833	0.6%	0.2%	1.4%	2.2%
SUFFOLK	2 911	1 042	7 159	11 112	2.3%	0.8%	5.6%	8.7%
WESTCHESTER	2 711	1 230	8 271	12 212	2.1%	1.0%	6.5%	9.6%
总计	35 986	10 175	81 449	127 610	28.2%	8.0%	63.8%	100.0%

将交叉表数据以一种更直观的方式表达出来也是可能的，然而，由于呈现了许多的数据，一般人很难领会复杂的图形。图 5-6 显示了该表所列计数的曲面图，该曲面图看起来有点像丘陵地带，计数是丘陵的高度，郡县沿一个边前进，渠道构成了第三维。这一曲面图显示出曼哈顿（属于纽约县）的“其他”渠道很高。尽管在本例中不是问题，但曲面图的山峰可能遮掩其他的丘陵和山谷。

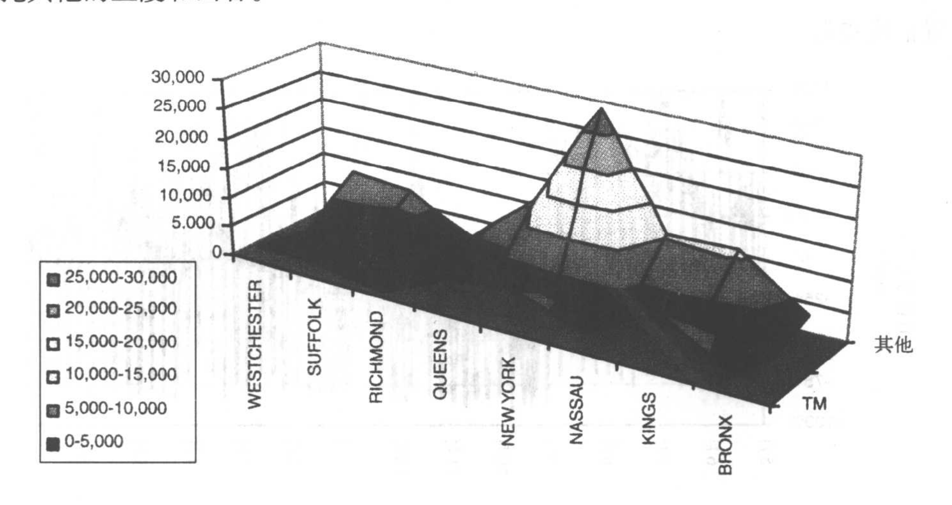


图 5-6 曲面图提供了交叉表数据的直观界面

5.2.2 观察连续变量

统计学最初是为了理解科学家收集的数据，大部分采用连续测量的形式。在数据挖掘中，我们不常遇见连续数据，因为还有很多描述性的数据。本节从描述统计学的观点讨论连续数据。

1. 连续变量的统计学度量

最基本的统计学度量描述一组数据只用单一值，最常用的统计量是均值或平均值（所有数值之和除以数据的个数）。需要说明的其他一些重要的概念是：

变动范围（range）：变动范围是样本中最小值和最大观察值之差，变动范围经常连同最小值和最大值一起观察。

均值（mean）：这也就是通常所说的平均值。

中值（median）：中值是把观察资料分为两个相等大小的组，一个组具有的观察资料比中值小，另一个组包含的观察资料比中值大。

众数（mode）：这是指最常出现的那个值。

中值用于一些不可能计算均值的情况，例如，当收入以 10 000 美元为界限报告，而最后一个类为“100 000 美元以上”时，每一组中的观察对象的个数是已知的，但实际数值是未知的。此外，中值很少被一些与其他数值相差很大的观察资料所影响。例如，如果比尔·盖茨迁居到你的街区，邻居的平均净资产将显著增加，但中值净资产可能根本不改变。

另外，各种各样的用于表征变动范围的方式也都是有用的，变动范围本身介于最小值和最大值之间。查看百分点信息常常是值得的，像查看第 25 个和第 75 个百分点，就可以了解

中间一半数值的极限。

图 5-7 显示了描绘按日订购数量变动范围和平均值的一个图表。该图表的垂直轴采用对数（log）标度，因为最小订购量在 10 美元以下而最大订购量超过 1 000 美元。事实上，最小值一直在 10 美元左右，平均值在 70 美元左右，最大值在 1 000 美元。与离散变量一样，对连续数值，利用时间图来掌握非预期事物何时出现也是有价值的。

2. 离差与标准差

离差度量样本的分散程度或者观测值围绕平均值聚集的紧密程度。变动范围不能很好地反应分散状况，因为它只考虑了两个极端值。去掉一个极值有时会显著改变变动范围。相反，离差考虑每一个值。某个观测值和样本均值两者之差称为偏差，离差被定义为偏差的平方的平均值。

标准差，即离差的平方根，最常用于度量分散的程度。它比离差更方便，因为它具有与观测值相同的单位，而不是其单位的平方。这就容许标准差本身可以用作度量单位。我们从前用过的 z 得分，就是用标准差测定的观测值与平均值的距离。利用正态分布，z 得分能够转换为概率或置信度。

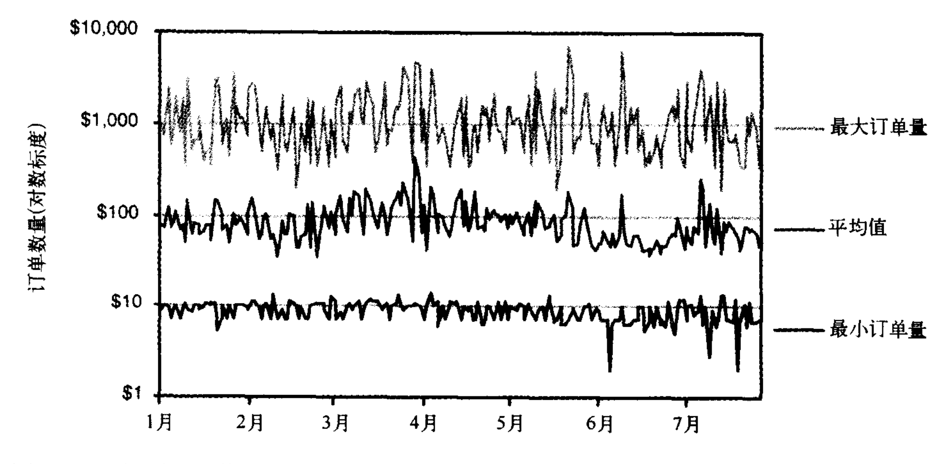


图 5-7　时间图表也能用于表示连续数值；此表显示了每日订购数量的变动范围和平均值

5.2.3　另一对统计概念

相关性（correlation）是考察一个变量的改变与另一变量的改变关联程度大小的度量。相关性的变动范围从 -1 到 1。相关性为 0 意味着这两个变量不相关。相关性为 1 意味着当第一个变量改变时，第二个肯定将按同一方向改变，尽管未必改变相同的数量。另一个相关性度量是 R^2 值，该值为相关性的平方，从 0（不相关）到 1（完全相关）。例如，圆的半径和周长完全相关，尽管后者比前者增长快得多。相关性为负值意味着两变量按相反方向改变，例如，海拔高度负相关于大气压力。

回归（regression）是用一对相关变量的一个值来预测另一个值的过程。回归的最普通形式是线性回归，这样叫是因为企图做出一条直线穿过样本中观测的 X 和 Y 对。一旦这条线被确定，就能够用于预测给定任意 X 值时的 Y 值，以及给定任意 Y 值时的 X 值。

5.3 测定响应

本节考察市场活动环境下的统计学思想。支持者－挑战者营销方法尝试与正常的业务不同的观念。例如，假定一个公司每个月派送一百万份促销插页来诱导客户。他们决定用一种促销插页的方法，即*支持者促销*，另一种促销是针对支持者促销的*挑战者*。比较二者的方法是：

——向900 000顾客派送支持者促销。

——向100 000顾客派送挑战者促销。

——决定哪一种更好。

问题是，如何知道一种促销比另一种好？本节引入置信度思想来更加详细地探讨这个问题。

5.3.1 比例标准误差

回答这一问题的方法是使用置信度区间的概念。在上述情况中，挑战者促销是向顾客的随机子集派送的。基于这一子集中的响应，这种促销对整个总体预期的响应情况是什么？

例如，假设初始群体中50 000人有可能对挑战者促销做出响应，而期望大约有5 000人做出实际响应，占收到挑战者促销总数的10%。如果正好这么多数量的人做出响应，那么该样本的响应率和总体的响应率均为5%。然而，有可能（尽管非常非常不太可能）所有50 000响应者都处于“收到该挑战者促销”的样本中，这将产生50%的响应率。另一方面，也有可能（并且也非常非常不太可能）50 000人中没有一个被选入样本中，那样响应率就是0%。在任何总体的10%的样本中，观测到的响应率可能低到0%，或者高达50%，当然这些是极端值，实际值可能更接近5%。

到目前为止，这一例子已展示了能够从总体中抽取许多不同的样本。现在让我们回顾这一情形并假定已经观察到样本中有5 000个响应者，那么关于整个人口总体，它告诉了我们什么？同样，有可能这些是总体中所有的响应者，因此低端估计为0.5%；另一种可能是，另外的每个人同样都是响应者（我们在选择样本时非常非常不幸），那么高端值将是90.5%。

这就是说，有100%的置信度说明，总体的实际响应率在0.5%到90.5%之间。有高的置信度是好的，然而，范围太宽没有用处。我们宁愿设置一个较低的置信度水平，通常，95%或99%的置信度对于市场营销目的就已经足够了。

响应值的分布遵循二项式分布。幸运的是，二项式分布与我们处理超过几百人的一个总体人群时的正态分布很相似。在图5-8中，锯齿状线条为二项式分布，平滑线条为相应的正态分布，它们几乎相同。

要解决的问题是：在假定大小为100 000的样本响应率为5%的条件下，决定相应的正态分布。像先前提到的那样，正态分布有两个参数，均值和标准差。均值是样本中观测到的平均值（5%）。要计算标准差，我们需要一个公式，统计学家已经给出了标准差（严格说来，这里是标准误差，但对我们来说，二者是等价的）和均值以及比例样本大小之间的关系，这叫做比例标准误差（SEP），公式为：

$$\mathrm{SEP}=\sqrt{\frac{p\times(1-p)}{N}}$$

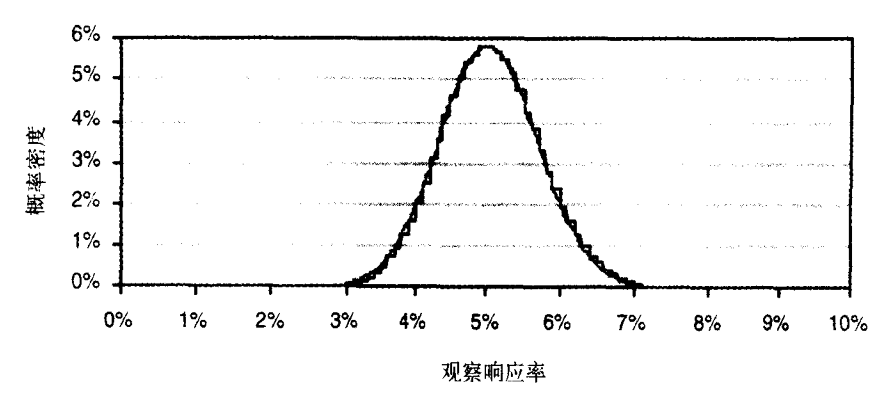


图 5-8 统计学家已证明，总体的实际响应率非常接近正态分布，其均值为观察到的样本响应，其标准差为比例标准误差（SEP）

在这个公式中，p 是平均值，N 是总体的大小。因此，与正态分布对应的有一个标准差，它等于观测响应与 1 减观测响应的乘积除以样本总数，然后取平方根。

我们已注意到，遵循正态分布的数据大约有 68% 位于一个标准差范围之内。对于 100 000的样本大小，该公式为 SQRT（5% * 95%/100 000），得数约为 0.07%。因此有 68%的置信度可以说，实际响应率在 4.93%和 5.07%之间。又注意到，在两个标准差之内的值略高于 95%，因此在 4.86%到 5.14%的范围内，有稍稍超过 95%的置信度。如果对该挑战者促销观测到 5%的响应率，那么我们有超过 95%的置信度认为，整个总体的响应率将在 4.86%和 5.14%之间。需要注意的是，这一结论成立的条件是“得到挑战者促销的人确实是从整个总体中随机选择出来的”这一事实。

5.3.2 使用置信界限比较结果

前一节讨论了运用于收到挑战者促销的一个群组的响应率的置信区间问题。在该案例中，实际上有两个响应率，一个是针对支持者，另一个是针对挑战者。这些响应率不同吗？注意观测到的比率也许是不同的（比方说 5%和 5.001%），但也许很难把它们互相区分开来。回答这一问题的一种途径是观察每一响应率的置信区间，看它们是否重叠。如果该区间并不重叠，那么响应率是不同的。

这个例子研究了支持者模型在 4.5%到 5.5%范围的响应率。单一的响应率在实践中可能是已知的。然而，研究一个范围内的响应率可以了解到，当响应率从低得多（4.5%）到相同（5.0%）再到大得多（5.5%）时，会出现什么情况。

95%的置信度是偏离均值 1.96 个标准差，因此最低值为均值减去标准差的这些倍数，最高值是均值加上这个值。表 5-2 显示了支持者模型从 4.5%到 5.5%的响应率范围的最低和最高界限。

表 5-2 支持者群组的 95%置信区间范围

响 应	大 小	SEP	95%的置信度	95%的置信度 * SEP	最低值	最高值
4.5%	900 000	0.0219%	1.96	0.0219% * 1.96=0.0429%	4.46%	4.54%
4.6%	900 000	0.0221%	1.96	0.0221% * 1.96=0.0433%	4.56%	4.64%
4.7%	900 000	0.0223%	1.96	0.0223% * 1.96=0.0437%	4.66%	4.74%

（续）

响 应	大 小	SEP	95%的置信度	95%的置信度 * SEP	最低值	最高值
4.8%	900 000	0.0225%	1.96	0.0225% * 1.96=0.0441%	4.76%	4.84%
4.9%	900 000	0.0228%	1.96	0.0228% * 1.96=0.0447%	4.86%	4.94%
5.0%	900 000	0.0230%	1.96	0.0230% * 1.96=0.0451%	4.95%	5.05%
5.1%	900 000	0.0232%	1.96	0.0232% * 1.96=0.0455%	5.05%	5.15%
5.2%	900 000	0.0234%	1.96	0.0234% * 1.96=0.0459%	5.15%	5.25%
5.3%	900 000	0.0236%	1.96	0.0236% * 1.96=0.0463%	5.25%	5.35%
5.4%	900 000	0.0238%	1.96	0.0238% * 1.96=0.0466%	5.35%	5.45%
5.5%	900 000	0.0240%	1.96	0.0240% * 1.96=0.0470%	5.45%	5.55%

注：响应率从4.5%变化到5.5%。95%的置信层次的界限用偏离均值1.96个标准差计算。

基于这些可能的响应率，可以指出该置信界限是否重叠。挑战者模型95%的置信界限是从大约4.86%到5.14%，当响应率是4.9%、5.0%或5.1%时，这些界限与支持者模型的置信界限重叠。例如，4.9%响应率的置信区间从4.86%到4.94%，它确实重叠于4.86%至5.14%。利用界限重叠方法，可以认为它们从统计学角度上说是相同的。

5.3.3 使用比例差值比较结果

界限重叠方法很容易，但其结果有点悲观。换句话说，即使置信区间重叠，我们也许仍然确信，该差异并不归因于给定置信层次下的偶然性。另一种方法是观察响应率之间的*差别*，而不是响应率本身。正像比例标准误差有公式一样，比例差值的标准误差（standard error of a difference of proportion，SEDP）也有一个公式：

$$\text{SEDP}=\sqrt{\frac{p1*\ (1-p1)}{N1+p2*\dfrac{(1-p2)}{N2}}}$$

这个公式看起来很像比例标准误差公式，只不过平方根内的部分是对每一组都重复类似计算。表5-3显示了这个公式应用于支持者－挑战者问题，支持者群组响应率从4.5%变化到5.5%时的情况。

表5-3 支持者和挑战者群组间差值95%的置信区间界限

挑战者		支持者		差值			
响应	大小	响应	大小	值	SEDP	z值	p值
5.0%	100 000	4.5%	900 000	0.5%	0.07%	6.9	0.0%
5.0%	100 000	4.6%	900 000	0.4%	0.07%	5.5	0.0%
5.0%	100 000	4.7%	900 000	0.3%	0.07%	4.1	0.0%
5.0%	100 000	4.8%	900 000	0.2%	0.07%	2.8	0.6%
5.0%	100 000	4.9%	900 000	0.1%	0.07%	1.4	16.8%
5.0%	100 000	5.0%	900 000	0.0%	0.07%	0.0	100.0%
5.0%	100 000	5.1%	900 000	−0.1%	0.07%	−1.4	16.9%
5.0%	100 000	5.2%	900 000	−0.2%	0.07%	−2.7	0.6%
5.0%	100 000	5.3%	900 000	−0.3%	0.07%	−4.1	0.0%
5.0%	100 000	5.4%	900 000	−0.4%	0.07%	−5.5	0.0%
5.0%	100 000	5.5%	900 000	−0.5%	0.07%	−6.9	0.0%

通过比例的差值，支持者的三种响应率拥有低于95%的置信度（换句话说，该p值超过5%）。如果挑战者响应率是5%，而支持者是5.1%，那么响应率的差值可能归因于偶然性。但是，如果支持者拥有5.2%的响应率，那么把这个差值归因于偶然性的可能将下降到不足1个百分点。

警告：置信区间只衡量抽样影响结果的可能性大小。也许我们需要考虑许多其他因素，以确定两个促销之间是否有显著的差别。要使比例差值方法真正起作用，每个群组必须完全随机地从整个总体选取。

5.3.4 样本大小

比例标准误差和比例差值的标准误差公式中都含有样本大小这一项。样本大小和置信区间大小之间是倒数关系：样本大小越大，置信区间就越狭窄。因此，如果想得到更多的置信度，就要使用更大的样本。

表5-4显示了不同挑战者群组大小的置信区间，假定挑战者响应率看做5%。对于很小的差别，置信区间很宽，经常的情况是置信区间太宽以至于没有用处。先前，我们已经说过正态分布是实际响应率的估计近似值；对于小的样本，该估计并不准确。统计学有几种方法来处理如此小的样本问题，但这通常并不能引起数据挖掘者多大兴趣，因为我们使用的样本要大得多。

表5-4 不同大小的挑战者群组95%的置信区间

响应	大小	比例的标准误差	95%的置信度	低点	高点	宽度
5.0%	1 000	0.6892%	1.96	3.65%	6.35%	2.70%
5.0%	5 000	0.3082%	1.96	4.40%	5.60%	1.21%
5.0%	10 000	0.2179%	1.96	4.57%	5.43%	0.85%
5.0%	20 000	0.1541%	1.96	4.70%	5.30%	0.60%
5.0%	40 000	0.1090%	1.96	4.79%	5.21%	0.43%
5.0%	60 000	0.0890%	1.96	4.83%	5.17%	0.35%
5.0%	80 000	0.0771%	1.96	4.85%	5.15%	0.30%
5.0%	100 000	0.0689%	1.96	4.86%	5.14%	0.27%
5.0%	120 000	0.0629%	1.96	4.88%	5.12%	0.25%
5.0%	140 000	0.0582%	1.96	4.89%	5.11%	0.23%
5.0%	160 000	0.0545%	1.96	4.89%	5.11%	0.21%
5.0%	180 000	0.0514%	1.96	4.90%	5.10%	0.20%
5.0%	200 000	0.0487%	1.96	4.90%	5.10%	0.19%
5.0%	500 000	0.0308%	1.96	4.94%	5.06%	0.12%
5.0%	1 000 000	0.0218%	1.96	4.96%	5.04%	0.09%

5.3.5 置信区间的真正含义

置信区间只是对结果的统计离差的一种度量。假设其他任何条件保持相同，它所测量的就是通过抽样过程引入的不精确量。它同时假定抽样过程本身是随机进行的——换句话说，一百万顾客中任何一个都有相等的可能性会被给予该挑战者促销，随机就是要随机。下面举

出的这些例子都是不应该出现的:

- 用加利福尼亚的顾客测试挑战者，其他任何人测试支持者;
- 使用最低的5%和最高的5%有价值的顾客测试挑战者，其他任何人测试支持者;
- 使用10%新近客户测试挑战者，其他任何人测试支持者;
- 使用有电话号码的顾客测试电话销售活动，其他任何人测试直接邮寄活动。

所有这些都是将总体分解为群组时会出现的偏离方式。前述所讨论的结果都假定没有这样的系统偏离 (bias)。当有系统偏离时，置信区间的公式就不正确。

使用置信区间的公式意味着在确定特定顾客是否收到支持者或挑战者信息时没有系统偏离。比如，假定有一个支持者模型用于预测顾客对支持者促销做出响应的可能性，一旦使用了该模型，那么挑战者样本将不再是一个随机样本，它将由支持者模型的剩余顾客组成。这样就引入了另一种形式的偏离。

另外一种情况是，挑战者模型也许只对特定市场或特定产品的顾客可用，这也引入了其他形式的偏离。在这种情况下，这些顾客应当与那些具有相同限制条件的、收到该支持者促销的顾客集进行比较。

另一种形式的系统偏离可能来自响应的方法。挑战者也许只通过电话接受响应，但支持者也许通过电话或者网络接受它们。在此情况下，挑战者响应也许会因为缺少网络这一渠道而变得低迷，或许需要对入站电话服务生进行特别训练以处理该挑战者促销。在某些特殊情况下，这可能意味着等待更长时间，又会造成了另一种形式的系统偏离。

置信区间仅仅是关于统计学和离差的说明，它不代表可能影响结果的所有其他形式的系统偏离，这些形式的偏离对结果的影响常常比样本方差更重要。下一节会讨论在市场营销中建立一个测试和对照实验，将这些问题引向细致深入的讨论。

5.3.6 实验的测试群组和对照群组大小

支持者-挑战者模型是一个双向测试的例子，它采用一个新方法（挑战者）与通常的商业活动（支持者）相对比。本节将讨论的问题是，对于当前目的如何确保测试群组（test group）和对照群组（control group）足够大。上一节讨论了如何确定样本响应率的置信区间，这里我们将这一逻辑反过来看，不是从群组的大小开始，而是从实验设计的观点考虑大小。这需要几项信息:

- 对其中一个群组估计响应率，称它为 p;
- 在响应率中我们期望慎重对待的差异（测试的敏锐度），称它为 d;
- 置信区间（比方说95%）。

这提供了确定测试群组和对照群组需要的样本大小的足够信息。例如，假定正常的商务有5%的响应率，我们期望以95%的置信度测量0.2%的差异，这意味着如果测试群组的响应率大于5.2%，那么该实验能够有95%的置信度检测到这个差值。

对于这种类型的问题，第一步是确定SEDP的值。也就是，如果我们愿意在95%的置信度下接受0.2%的差异，那么对应的标准误差是多少？95%的置信度意味着偏离均值1.96个标准差，因此答案就是将该差值除以1.96，得到0.102%。一般地说，该过程是把p值（95%）转换为z值（这可以用Excel函数NORMSINV完成），然后将期望的置信度除以这个值。

下一步是将这些值代入 SEDP 公式中，为此，我们假设测试群组和对照群组具有相同大小：

$$\frac{0.2\%}{1.96}\sqrt{\frac{p*(1-p)}{N+(p+d)}*\frac{(1-p-d)}{N}}$$

将刚才描述的值（p 为 5%，d 为 0.2%）代入后的结果是：

$$0.102\%=\sqrt{\frac{5\%*95\%}{N}+\frac{5.2\%*94.8\%}{N}}=\sqrt{\frac{0.0963}{N}}$$

$$N=\frac{0.0963}{(0.00102)^2}=66875$$

因此，两个同样拥有 92 561 个样本的群组，可以用于以 95% 的准确度测量响应率中出现的 0.2% 的差值。当然，这并不是保证结果将最少差 0.2 个百分点，只是说对于最少具有这样大小的对照组和测试群组，在响应率中若出现 0.2% 的差异应该能够测量到，并有显著的统计学差异。

测试群组和对照群组的大小会影响到如何解释结果，然而这种影响可以在测试之前被提前确定。在进行测试之前确定测试群组和对照群组的敏锐程度是值得的，这样可以确保该测试能产生有用的结果。

提示：在进行一个市场测试之前，应该通过计算响应率差值确定测试的敏锐程度，而且响应率差值计算过程中要设置较高的置信度（例如 95%）。

5.4 多重比较

到目前为止，讨论只用了一种对比的例子，例如两个总统候选人或测试群组和对照群组之间的差值。我们常常同时运行多个测试：例如，可能试验三种不同的挑战者信息，以决定其中之一是否比通常的营销信息产生更好的结果。因为处理多重测试确实影响基础统计数字，所以理解发生了什么是重要的。

5.4.1 多重比较下的置信层次

设想有两个群组已被测试，并且获知两组响应差值有 95% 归因于抽样差异之外的因素，那么一个合理的结论是在两个群组之间确实存在差异。在一个精心设计的测试中，最可能的原因有可能是信息、服务或待遇等方面的差别。

Occam 的剃刀学说告诉我们，应当尽可能采用最简单的解释，不要添加额外的东西。对于响应率差异的最简单假说是“该差异并不重要”，这些响应率实际上近似于相同的数值。如果该差异是重要的，那么我们需要寻找导致出现这种差异的理由。

现在考虑相同的情形，但不同的是获知实际上有 20 组正被测试，展示的只是其中的一对。现在可能得到一个非常不同的结论。如果 20 组正被测试，那么应该期望它们中有一个会超过 95% 的置信界限，出现的原因纯粹应归因于可能性，因为 95% 意味着 19/20。你不能再推断该差值归因于测试参数；相反，很可能该差异应归因于抽样差异，这是最简单的假说。

置信层次只是基于单一比较。当有多重比较时，前提条件就不正确，因此前面所计算的置信度就不太充分了。

5.4.2 Bonferroni 修正

幸运的是，意大利数学家 Carlo Bonferroni 提出的简单修正法可以校正这个问题。我们一直在观察置信度问题，即前面所说的某些值有 95% 的机会出现在 A 和 B 之间。为此考虑以下几种情况：

- 有 95% 的概率 X 处于 A 和 B 之间；
- 有 95% 的概率 Y 处于 C 和 D 之间。

Bonferroni 希望知道这两者都为真的概率大小。观察它的另一种方式是确定一个或者另一个为假的概率，这样计算更容易些。如果第一个为假的概率是 5%，另一个为假的概率也是如此，则两者任意为假的概率是它们的和（即 10%）减去二者同时为假的概率（0.25%）。因此，两个命题同时都为真的概率约为 90%。

从 p 值来观察，两个命题合起来的 p 值（10%）近似于两个单独命题的 p 值之和。这不是偶然的，实际上，把任何数目命题的 p 值作为每一个命题的 p 值之和来计算是合理的。如果有 8 个变量具备 95% 的置信度，那么在任意给定时间，我们预期 8 个变量将有 60% 的可能同时出现在范围内（因为 8 * 5% 为 40% 的 p 值）。

Bonferroni 反过来应用了这一观测资料。如果有 8 个测试并且希望总体是 95% 的置信度，那么 p 值的限度需为 5%/8 = 0.625%。换句话说，每一观测资料需要至少 99.375% 的置信度。Bonferroni 修正就是按照做出比较的数目分配期望的 p 值界限，以便得到所有比较的 1 − p 的置信度。

5.5 卡方检验

比例差值方法对于估计活动有效性及其他相似情形是一个强有力的方法。不过，还有另一个统计测试方法可以使用，这就是卡方检验（chi-square test），它是特别为多重测试且至少有两个离散结果（例如响应和非响应）的情形设计的。

卡方检验的吸引力在于它非常适合于多重测试群组和多重结果，只要不同的群组相互截然不同。实际上这几乎是使用这一测试时惟一的重要规则。正如下一章关于决策树所描述的那样，卡方检验是决策树最初形式之一的基础。

5.5.1 期望值

开始卡方计算需要在一个表格中排布数据，如表 5-5 所示。这是一个简单的 2×2 表格，代表在有两种结果（比方说响应或非响应）的测试中的测试群组和对照群组。表中也显示了每一列和行的合计值，亦即，响应者和非响应者（每列）的总数，以及在测试群组和对照群组（每行）中的总数目。响应列被添加上去仅用作参考，并不是计算的一个组成部分。

如果在这些群组之间，数据以一种完全没有偏离的方式被分裂成两半，情况会怎样？就是说，如果在表格中的行列之间真的没有差别，结果会怎样？这是一个十分合理的问题。假定响应者和非响应者的数量相同，并假定支持者和挑战者群组的大小相同，我们可以计算出期望值。更确切地说，我们能够计算每一个单元格中的期望值，假定行和列的大小与原始数据相同。

计算期望值的一种方法是：通过计算下列四个量中每个量的值，计算每一列中每一行的

比例，像在表 5-6 中显示的那样：

- 做出响应的人的比例
- 没有响应的人的比例

然后这些比例被乘以每一行的计数以获得期望值。当表中的数据有更多列或更多行时，这种计算期望值的方法就可以进行。

表 5-5　为进行卡方检验部署支持者－挑战者数据

	响应者	非响应者	合　计	响应率
支持者	43 200	856 800	900 000	4.80%
挑战者	5 000	95 000	100 000	5.00%
合计	48 200	951 800	1 000 000	4.82%

表 5-6　对表 5-5 中的数据计算期望值和预期离差

	实际响应			期望响应		离　差	
	是	否	合计	是	否	是	否
支持者	43 200	856 800	900 000	43 380	856 620	-180	180
挑战者	5 000	95 000	100 000	4 820	95 180	180	-180
合计	48 200	951 800	1 000 000	48 200	951 800		
总体比例	4.82%	95.18%					

期望值是很有意思的，因为它显示了如果没有其他因素影响，数据会如何分解。请注意期望值的单位与每一单元格的单位是相同的，通常是顾客计数，因此它实际上具有一定的意义。同样地，期望值之和与原始表格中的所有单元格之和相同。该表也包含了离差，即观测值和期望值之间的差值。在这种情况下，离差都具有相同的值，但是有不同的符号，这是因为原始数据有两行和两列。在本章的稍后部分，有一个使用了更大表格的例子，在这个表格中各个离差值是不同的，可是，每一行和每一列的离差总是会相互抵消，因此每行中的离差之和总是为 0。

5.5.2　卡方值

离差是观察数值的好工具，但它并不提供关于离差是预期的或非预期的信息。要想做到这点需要使用更多统计学工具，这就是由英国统计学家 Karl Pearson 在 1900 年提出的卡方分布。

每一单元格的卡方值（chi-square）可以由下式简单计算：

$$\text{Chi-square}\ (x)\ =\sqrt{\frac{(x-\text{expected}\ (x))^2}{\text{expected}\ (x)}}$$

整个表的卡方值是表中所有单元格的卡方值之和。注意，卡方值总是为 0 或者正数。同样，当表中的数值与期望值（expected）相符时，则总的卡方值为 0。这是我们能够做到的最好程度了，当偏离期望值的离差增大时，卡方值也随之增大。

可惜的是，卡方值不遵循正态分布。这实际上是很明显的，因为卡方值总是正数，而正态分布是对称的。值得庆幸的是，卡方值符合另外一种分布，这种分布同样是我们熟知的。可是，卡方分布不仅依赖于其数值本身，而且依赖于表格的大小。图 5-9 显示的是几个卡方

分布的密度函数。

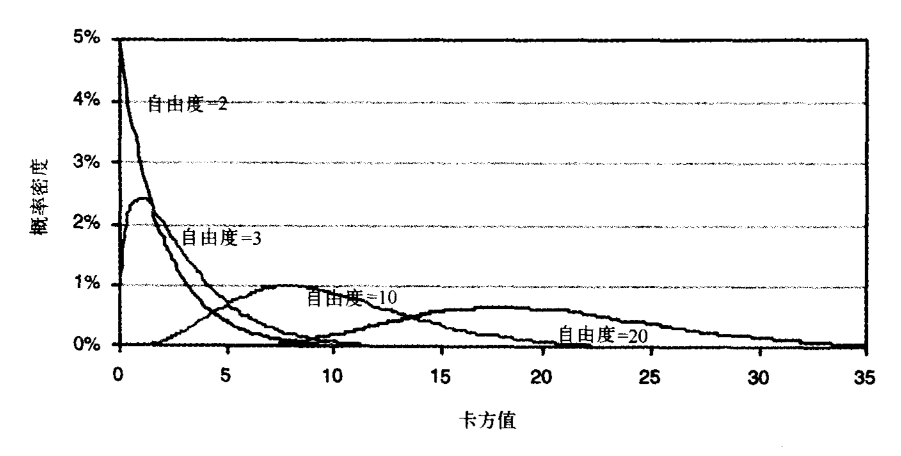


图 5-9 卡方分布依赖于所谓的自由度。但一般而言，它起始较低，峰值超前，逐步下降

卡方依赖的是自由度。跟概率论和统计学中的许多概念不同，自由度很容易计算，但其意义解释相对困难一些。某个表格的自由度等于行数和列数各自减一后相乘，在前一例子中的 2×2 表格有 1 个自由度；一个 5×7 的表格会有 24（4×6）个自由度。后面的“自由度”部分对这一问题进行更详细的讨论。

警告：在任何单元格中，当期望值数字小于 5（我们宁愿推荐一个稍微高一点的界限）的时候，卡方检验不起作用。

尽管对于大型数据挖掘这并不是什么问题，但当分析小型测试结果时它可能就成为问题了。

使用卡方检验的步骤是：

- 计算期望值；
- 计算偏离期望值的离差；
- 计算卡方（离差的平方除以期望值）；
- 对表格的全部卡方值求和；
- 计算观测值归因于偶然性的概率（在 Excel 中，可以使用 CHIDIST 函数）。

自 由 度

自由度（dof）所指的意思是欲描述期望值的表格需要多少不同的变量。这是对表格中数据受多大程度约束的一个度量。

如果该表格有 r 行 c 列，那么表中就有 $r \times c$ 个单元格。假如表格中没有加入约束条件，这就是需要的变量数目。然而期望值的计算还是要加入一些约束条件的。尤其是，对于原始表格来说，每一行数值之和与期望值之和是相同的，因为每一行的合计是固定的。换句话说，如果一个值缺失了，利用该约束条件，可以从整行合计中减去该行其余数值之和从而重新计算它。这表明该自由度为 $r \times c - r$。对于列也存在同样的情形，这样就会产生自由度为 $r \times c - r - c$ 的估算。

然而，另外有一个约束条件：所有行合计的总和与所有列合计的总和必定相同，所以我们实际上多计算了一个约束条件，因此该自由度实际为 $r \times c - r - c + 1$。换另一种方式，该

式可写作 $(r-1)\times(c-1)$。

结果可能是，表格中的数值分布是由于随机波动而不是一些外部因素的影响造成的。就像Occam的剃刀学说所提出的，最简单的解释是各种因素根本没有造成差别，观测值与期望值之间的差值完全在预期范围之内。

5.5.3 卡方与比例差值的比较

卡方与比例差值可应用于同样的问题。尽管所得的结果不完全相同，但结果足够相似，已令人满意。在前面的表5-4中，我们使用比例差值方法，针对一系列支持者响应率，确定出支持者和挑战者产生相同结果的可能性。在表5-7中，使用卡方计算代替比例差值法对这个问题重新进行计算，由卡方检验得到的结果与从比例差值得到的结果十分相似——这是考虑这两种方法有多大差别时一个值得注意的结果。

表5-7 对于表5-4中比例差值的卡方计算

挑战者		支持者			挑战者期望值		支持者期望值		挑战者卡方		支持者卡方		卡方		比例差值
响应	无响应	响应	无响应	总响应率	响应	无响应	响应	无响应	响应	无响应	响应	无响应	数值	p值	p值
5 000	95 000	40 500	859 500	4.55%	4 550	95 450	40 950	859 050	44.51	2.12	4.95	0.24	51.81	0.00%	0.00%
5 000	95 000	41 400	858 600	4.64%	4 640	95 360	41 760	858 240	27.93	1.36	3.10	0.15	32.54	0.00%	0.00%
5 000	95 000	42 300	857 700	4.73%	4 730	95 270	42 570	857 430	15.41	0.77	1.71	0.09	17.97	0.00%	0.00%
5 000	95 000	43 200	856 800	4.82%	4 820	95 180	43 380	856 620	6.72	0.34	0.75	0.04	7.85	0.51%	0.58%
5 000	95 000	44 100	855 900	4.91%	4 910	95 090	44 190	855 810	1.65	0.09	0.18	0.01	1.93	16.50%	16.83%
5 000	95 000	45 000	855 000	5.00%	5 000	95 000	45 000	855 000	0.00	0.00	0.00	0.00	0.00	100.00%	100.00%
5 000	95 000	45 900	854 100	5.09%	5 090	94 910	45 810	854 190	1.59	0.09	0.18	0.01	1.86	17.23%	16.91%
5 000	95 000	46 800	853 200	5.18%	5 180	94 820	46 620	853 380	6.25	0.34	0.69	0.04	7.33	0.68%	0.60%
5 000	95 000	47 700	852 300	5.27%	5 270	94 730	47 430	852 570	13.83	0.77	1.54	0.09	16.23	0.01%	0.00%
5 000	95 000	48 600	851 400	5.36%	5 360	94 640	48 240	851 760	24.18	1.37	2.69	0.15	28.39	0.00%	0.00%
5 000	95 000	49 500	850 500	5.45%	5 450	94 550	49 050	850 950	37.16	2.14	4.13	0.24	43.66	0.00%	0.00%

5.6 示例：区域和起点的卡方

一家大型的面向消费者的公司曾在纽约地区进行过支持者获取方面的调查活动。这一分析的目的是观察他们的获取渠道，试图增加对该区域内不同部分的了解。针对这一分析目的，令人感兴趣的渠道有三种：

电话销售（telemarketing）：通过拨打销售电话获取的客户（注意：这一数据是在“全国禁止呼叫列表”生效前收集到的）；

直接邮寄（direct mail）：对直接邮寄有响应的顾客；

其他：通过其他方法进来的顾客。

这个令人感兴趣的区域由纽约州的八个郡组成，其中有五个是纽约市的行政区，另外两个（Nassau郡和Suffolk郡）在Long Island，还有一个（Westchester）位于城市的正北边。

这一数据已经出现在先前的表 5-1 中，分析的目的是确定按渠道和郡县的起始细目分类是否归因于偶然因素，或者是否还有其他的一些因素在起作用。

这一问题特别适合于卡方计算，因为数据可以被排布到行和列中，每个客户只会在一个单元格中被计数。表 5-8 显示了表中每一组合的离差、期望值和卡方值。注意：在这一示例中卡方值常常很大，该表的总体卡方得分是 7 200，相当大；归因于偶然性的总体得分概率基本为 0，即是说，分渠道和分区域的起始之间的差别不是由于样本差异引起的，还有其他因素在起作用。

表 5-8 对郡县和渠道进行卡方计算的示例

郡　县	期望值			离差			卡方		
	电话销售	直接邮寄	其他	电话销售	直接邮寄	其他	电话销售	直接邮寄	其他
BRONX	1 850.2	523.1	4 187.7	1 362	−110	−1 252	1 002.3	23.2	374.1
KINGS	6 257.9	1 769.4	14 163.7	3 515	−376	−3 139	1 974.5	80.1	695.6
NASSAU	4 251.1	1 202.0	9 621.8	−1 116	371	745	293.0	114.5	57.7
NEW YORK	11 005.3	3 111.7	24 908.9	−3 811	−245	4 056	1 319.9	19.2	660.5
QUEENS	5 245.2	1 483.1	11 871.7	1 021	−103	−918	198.7	7.2	70.9
RICHMOND	798.9	225.9	1 808.2	−15	51	−36	0.3	11.6	0.7
SUFFOLK	3 133.6	886.0	7 092.4	−223	156	67	15.8	27.5	0.6
WESTCHESTER	3 443.8	973.7	7 794.5	−733	256	477	155.9	67.4	29.1

下一步是确定哪些数值偏高、哪些数值偏低，以及具有多大的概率。它吸引我们使用该表的自由度，将每一个单元格中的卡方值转换成一个概率，该表是 8×3 的，因此它有 14 个自由度。然而，这并不是要做的恰当的事，卡方的结果是对整个表格的，把每一个得分转化成概率不会产生有效的结果，因为卡方得分不可累加。

另一种可选的方法被证明是更准确的，思路是将每一单元格与其他的任意一个相比较，结果给出有两列和两行的一个表格，如表 5-9 所示。其中一列是原始单元格列，另一列是其余全部列之和；一行是原始单元格的行，另一行是其余全部行之和。

表 5-9 对 Bronx 郡和电话销售的卡方计算

郡　县	期望值		离差		卡方	
	电话销售	非电话销售	电话销售	非电话销售	电话销售	非电话销售
BRONX	1 850.2	4 710.8	1 361.8	−1 361.8	1 002.3	393.7
非 BRONX	34 135.8	86 913.2	−1 361.8	1 361.8	54.3	21.3

其结果是一组 Bronx 郡与电话销售组合的卡方值，绘于有 1 个自由度的表格中。Bronx 郡－电话销售得分本身是对于一个 2×2 表格全部的卡方值的一个良好近似（这里假定原始单元格大约具有相同大小）。卡方值的计算使用这个值（1002.3）和 1 个自由度。方便的是，对这一单元格的卡方计算与该单元格的卡方原始计算相同（虽然其余的数值没有任何相匹配的计算），这样就不必进行额外的计算。

这就是说，每一种变量组合的效果评估可以使用单元格中的卡方值和 1 个自由度得到。其结果是包含一组 p 值的表格，其中某个给定的格是由偶然性引起的，如表 5-10 所示。

表 5-10 对每一郡县和渠道组合估计的 p 值，没有对比较作修正

郡　县	电话销售	直接邮寄	其　他
BRONX	0.00%	0.00%	0.00%
KINGS	0.00%	0.00%	0.00%
NASSAU	0.00%	0.00%	0.00%
NEW YORK	0.00%	0.00%	0.00%
QUEENS	0.00%	0.74%	0.00%
RICHMOND	59.79%	0.07%	39.45%
SUFFOLK	0.01%	0.00%	42.91%
WESTCHESTER	0.00%	0.00%	0.00%

然而，因为同时进行了许多比较，还需要做出二次修正，Bonferroni 对此进行了调整，他把每一个 p 值与相比对的数目（表格中单元格的数目）相乘。为了最终的表达目的，需要将 p 值转换为它的相对值置信度，通过乘以离差的符号以得到一个有正负之分的置信度。图 5-10 显示了这些结果。

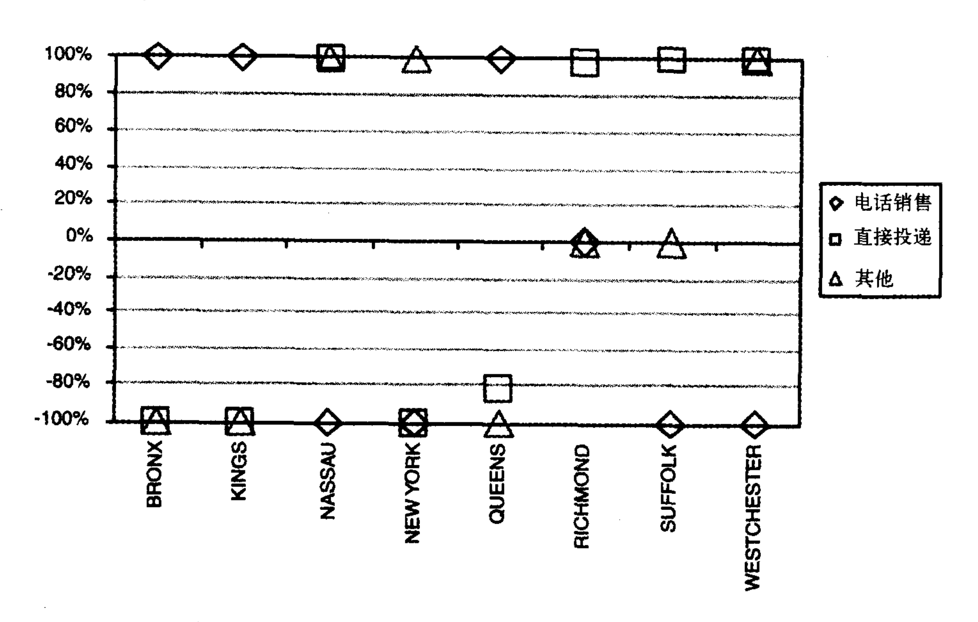


图 5-10　这个图表明对每个郡县和地区组合的有符号的置信度值。接近 100% 和 - 100% 的数值占优势，表明观测到的差异在统计学上是显著的

这一结果很令人关注。首先，几乎所有数值都接近 100% 或 - 100%，意味着在郡县之间有统计学上的显著差异。事实上，电话销售（菱形）和直接邮寄（方形）总是处于相反的两端，两者之间有直接的相反关系。在三个郡县（Manhattan、Nassau 和 Suffolk）直接邮寄高而电话销售低。在这些郡县有许多富人区，暗示与电话销售相比较，富有的顾客更可能响应直接邮寄；当然，这可能也意味着直接邮寄活动是面向这些区域的，而电话销售面向其他区域，于是这种地理分布差异是由商业运作造成的。要确定这些可能性中哪种是正确的，我们既需要知道哪些人被联系过，也需要知道哪些人做出了响应。

5.7 数据挖掘和统计学异同

接下来8章中讨论的许多数据挖掘技术是由统计学家发明的，或者现在已被集成到统计软件中，它们是标准统计学的延伸。尽管数据挖掘者和统计学家使用相似的技术来解决相似的问题，但数据挖掘方法在几个方面不同于标准统计方法：

- 数据挖掘者倾向于忽略原始数据中的测量误差；
- 数据挖掘者假定有足够多的数据和足够强的处理能力；
- 数据挖掘假定时时处处具有相关性；
- 在商业界设计试验可能很困难；
- 数据已被截取（truncated）或审查（censored）。

这些仅仅是方法上的差异，它们不是对立的。这些差异从某种程度上说明，数据挖掘者要处理的商业问题与激励统计学发展的科学问题是不同的。

5.7.1 原始数据中没有测量误差

统计学最初源于科学上对量的测量，诸如头骨的宽度或星星的亮度。这些测量是定量的，且精确的测量值依赖于诸如测量设备的类型和环境温度等因素。特别是，两人同时进行相同的测量将产生稍微不同的结果，该结果可能相差5%或者0.05%，但确实有差别。传统上，统计学把观测值看做落入置信区间。

另一方面，顾客去年1月份付款的数量非常好理解——可以精确到最后一分钱。顾客的定义也许有一点模糊，一月份的定义也许是模糊的（考虑5-4-4财务周期），但是付款数量是精确的，没有测量误差。

商业数据是有误差来源的。特别要关注的是操作系统误差，在所收集的数据中它能导致系统偏差。例如，时钟相位差也许意味着，本应按某一个序列发生的两个事件似乎有可能按另一个顺序发生；一个数据记录可能把星期二标记为更新日期，但它实际更新的时间是在星期一，因为该更新过程在午夜刚过就进行。这种形式的偏差是系统的，潜在地代表可能被数据挖掘算法拾取的虚假模式。

在商业数据和科学数据之间一个较大的区别是后者有许多连续值，而前者有许多离散值。甚至金钱的数量也是离散的（两个值可能只差几美分或某一类似量），即便是该数值能用实数表示。

5.7.2 有大量的数据

传统上，统计学被应用于短小的数据集（至多几千行），通常只有较少的列（少于12个），其目标是从数据中压榨出尽可能多的信息。在数据收集代价昂贵或费劲的领域——诸如市场调查、汽车碰撞试验或火星土壤化学成分试验中，这仍然是重要的。

相反，商业数据是非常庞大的。亟待解决的问题是了解正在发生的任何事情，而不是任何可能的事情。幸运的是，目前有足够的计算能力可处理如此巨大数量的数据。

抽样理论是统计学的一个重要部分。这部分内容可用于解释数据子集（样本）的结果与整体的关系。当计划进行一次民意测验时这是很重要的，因为不可能询问每个人问题；相反地，调查者是通过询问很小的样本来导出总体的看法。然而，当全部数据可用时，这点就很

不重要了。通常情况下，最好是使用所有可用的数据，而不是它的一个小子集。

有几种情况并非一定如此，一种可能是有太多的数据：无需在上千万的客户基础上建模，而代之以在几十万客户数据上建模——至少可以知道如何建立更好的模型。另一种可能的情况是得到一个没有代表性的样本：例如，这样的样本可能有相同数量的流失者和非流失者，尽管原始数据有不同的比例。然而，通常使用更多数据比样本裁减及使用更少数据要好，除非有抽样裁减的好理由。

5.7.3 时间从属性随处出现

几乎数据挖掘中的所有数据都具有与之相关联的时间从属性。顾客对营销工作的反应随时间改变，潜在顾客对竞争性服务的反应随时间改变，比较本年度与上一年市场营销活动的效果，很少会产生正好相同的结果。当然，我们也不期待出现相同的结果。

另一方面，我们确实希望科学试验产生相似的结果，不管该试验何时进行。科学定律被认为是永恒的，它们不随时间改变。与之形成对照的是，商业环境每天都在变化。统计学经常把重复的观测视为独立的观测，也就是说，一个观测与另一个并不相似。相反，数据挖掘必须经常考虑数据的时间成分。

5.7.4 试验是艰难的

数据挖掘不得不在现有商业实践的约束中进行研究。这使得编排试验变得困难，有以下几个主要原因：

- 商业界也许不愿意资助为了长期获益而减少短期收益的努力；
- 商业活动过程可能妨碍精心设计的试验方法的实施；
- 影响试验结果的可能因素也许并不明显；
- 时限扮演关键角色并可能致使结果没有用处。

上述这些因素中，前两项是最困难的。第一条只是说试验没有得到实施，或者是实施得如此拙劣以至于结果是无用的。第二条造成的问题是，一个看上去设计精良的试验可能没有正确执行，在计划一个试验时总会有些羁绊，有时候这些羁绊会使读懂结果变得不太可能。

5.7.5 数据审查和截取

用于数据挖掘的数据经常是不完善的，通常会以两种特殊方式出现。因为被测量的任何东西是不完整的，从而导致被审查的数值不完善。一个例子是顾客保有期：对于活跃的顾客，我们知道其保有期肯定大于当前的保有期，然而我们不知道哪些顾客明天将停止，哪些顾客将自现在起 10 年后停止。实际的保有期总是大于观测值，并且直到该顾客实际停止于将来的某一特定未知点才能知晓。

图 5-11 显示了另一个具有同样结果的情形。这一曲线显示了某零售商关于一种产品的销售和库存。销售总是小于或等于库存。可是，在标注了 X 的日子里，库存卖完了，那么这些天的潜在销售会是多少？潜在销售大于或等于观测的销售——这是被审查数据存在问题的另一个例子。

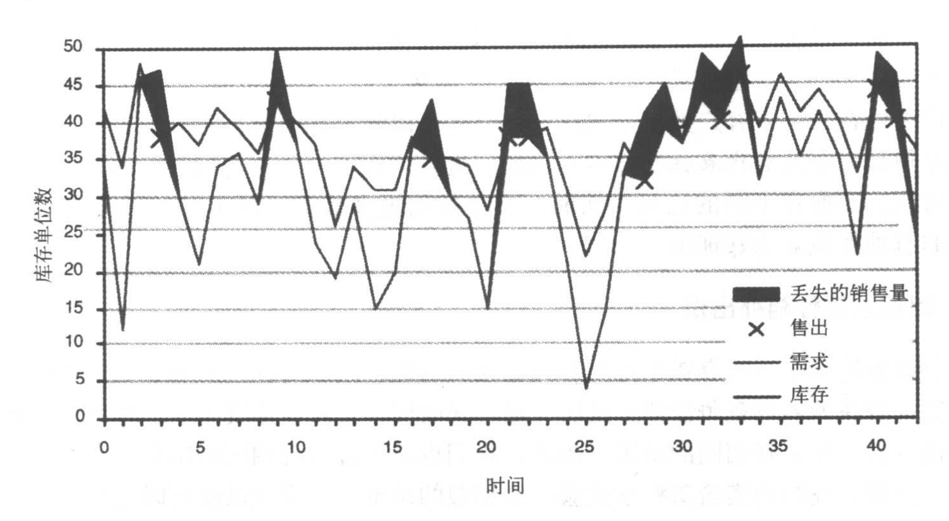


图 5-11 产品销售和库存的时间序列表明被审查数据存在的问题

被截取的数据在样本偏置方面造成了另一个问题：被截取的数据是不包含在数据库中的，经常是因为它太旧了。例如，当A公司收购B公司时，它们的系统被合并。来自B公司的活跃顾客常常被直接搬进A公司的数据仓库中。就是说，所有在给定日期活跃的顾客被挪走，前一天刚停止的顾客没有被挪走。这是一个左截取的例子，它在整个公司数据库会随处可见，通常并没有警告标识（除非文件非常好地表述了数据仓库中哪些存在及哪些不存在）。在考察客户关系什么时间开始时，这可能导致混乱——发现在合并前5年开始的所有客户都神秘地活跃了至少5年，当然这并非由于一个神奇的获取计划造成地，而是因为先前停止的所有那些人都被排除在外了。

5.8 小结

本章讨论了对于分析数据很有用的一些基本的统计学方法。在考察数据的时候，观察直方图和累积直方图，看哪些值最普通是非常有用的。尽管如此，更重要的是随时间考察数值。

统计学关注的重要问题之一是观测值是否是预期的。对于这点，偏离平均值的标准差数目（z得分）能够用于计算该值归因于偶然性的概率（p值）。高的p值意味着原假设为真，换句话说，没有出现任何有意义的事；低的p值暗示其他因素可能影响结果。可以依靠正态分布把z得分转换为p值。

商业问题经常需要分析表示为比例的数据。幸运的是，这些工作与正态分布很相似。比例标准误差公式（SEP）使得在诸如响应率这样的比例上可以定义置信区间。比例差值的标准误差（SEDP）使我们可以确定两个值是否相似，可以通过定义这两个值之间差值的置信区间来完成。

当设计营销试验时，SEP和SEDP都能够用于样本集大小测试和对照群组选择。尤其是，这些群组应当足够大，以便可以测量具有足够高置信度的响应率差值。对具有多于两个群组的测试，在设置群组大小时需要考虑某种调整，如Bonferroni的修正。

卡方检验是另一个常常很有用的统计方法。这一方法直接对排布成行和列的数据计算估

计值。基于这些估计值，卡方检验能够确定该结果是可能的还是不可能的。如同本章中的示例所示，卡方检验和 SEDP 方法得出相似的结果。

统计学家和数据挖掘者解答相似的问题。但是因为历史的差异和所面对问题本质的差别，在处理问题的方法上有一些差异。数据挖掘者通常面对许许多多具有很少测量误差的数据，这些数据随时间改变，且数值有时不完善。数据挖掘者必须对商业过程中引入数据的偏差保持特别警觉。

接下来的 8 章将研究建模和理解数据所需现代技术的更多细节，其中的许多技术已被统计学家采用，并以此为基础在这一领域应用了百年以上的时间。

第6章 决 策 树

决策树对于分类和预测是强有力的常用工具。基于树的方法之所以有吸引力，很大程度上是因为决策树代表着规则。规则可以很容易地用英语表达，以便我们能够理解，也能在数据存取语言中表示，比如用 SQL 在特定的类别中检索记录。决策树对于探测数据也很有用，可以了解从大量的候选输入变量到一个目标变量的关系。决策树把数据探查（data exploration）和建模结合在一起，即使建立最终模型时使用一些其他技术，它也是建模过程中强有力的第一个步骤。

在模型准确度（accuracy）和模型透明度之间时常要做些权衡。在某些应用中，分类或预测的准确度是惟一重要的事情，如果一个直接邮寄公司得到一个模型，能够准确地预测潜在顾客池中哪些成员最可能响应某个诱导，该公司也许不会关心该模型如何工作或为什么起作用。在另外一些情形下，阐述决策动机的能力则是至关重要的。例如，在保险业中，一些法律禁止基于某些变量的歧视。保险公司也许会发现自身处在这样的位置，他们不得不向法庭论证在允许或拒绝某个保险项目时没有使用非法的歧视性惯例。类似地，不管是信贷员还是贷款申请者，听到贷款申请是基于计算机产生的规则被拒绝的（例如，收入低于某一限额值并且现有周转账户超出另外某一限额），比听到该决定是由对其决定不提供任何解释的神经网络做出的更可以接受。

本章首先通过实例介绍什么是决策树、它们如何工作以及如何用于分类和预测问题，然后描述用于建立决策树的核心算法并讨论该核心算法的一些最流行的变体。作者精心选取的实例演示了决策树的效用和一般适用范围，说明实践中必须予以考虑的事项。

6.1 什么是决策树

决策树是一种结构。通过应用简单的决策规则，利用这种结构可以将大型记录集分割为相互连接的小记录集。通过每一次连续分割，结果集中的成员彼此变得越来越相似。18 世纪 30 年代，瑞典植物学家 Carl Linnaeus 发明了一种常见的生物分类方法，将生物划分为界、门、纲、目、科、属、种，这就是一个很好的例子。在动物界中，某一特定动物如果生有脊髓就被划分到脊椎动物门中；附加的特征用于将脊椎动物进一步细分为鸟、哺乳动物、爬行动物纲等；这些纲再进一步细分，直到分类学的最底层，同一个种的成员不仅在形态学上相似，而且能够繁殖产生后代。

决策树模型包含一系列规则，按照某个相关的特定目标变量，将大量包含不同种类的总体分割为小的、更相似的群组。决策树可以像 Linnaeus 以及后来的一代代分类学者所做的那样，通过手工方法辛苦地建立起来，也可以通过将某种决策树算法应用于包含预分类数据的模型组而自动产生，本章最关注的是自动产生决策树的算法。目标变量通常是分类属性，决策树模型可用于计算给定记录归属于某一个类别的概率，也可通过将记录分配到最可能的类来给记录分类。当然，决策树也能够用于估计连续变量的值，尽管其他技术更适合于这一任务。

6.1.1 分类

任何熟悉“二十问题”游戏的人将毫不费力地理解决策树是如何分类记录的。在该游戏中，一个玩家想出一个所有参与者可能知道或认识的特定的地点、人物或者事物，但是该玩家不给出关于其特性的任何提示，其他玩家通过一连串“是或否”的提问尝试发现它是什么。从“它是否比面包盒大?”这样的问题一直猜到“金门桥”，一个好玩家自始至终很少用满“20个问题”这一配额。

决策树代表了这样一系列连续的问题。正像游戏中那样，对第一个问题的回答决定了后续的提问，前面的问题先创建具有许多成员的宽泛范畴，后续问题将宽泛范畴分割为越来越小的集合。如果精心挑选所问的问题，那么也许只需几个问题就足以正确分类引入的记录。

“二十问题”游戏说明了用树来对记录追加分数或分类的过程。记录在根结点处进入树，根结点应用一个测试来确定该记录接下来将遇到哪个*子结点*。尽管有不同算法可用于选择初始测试，但目标总是相同的：选择在目标分类中最能判别的测试；这一过程反复进行，直到记录到达*叶结点*为止。所有终结于该树某一个给定叶的记录在分类上的路线是相同的，从根到每个叶只有惟一的路径，那个路径就是一个用于分类记录*规则*的表达式。

不同的叶可能产生相同的分类，尽管每个叶给出的分类可能依据不同的理由。例如，在一个按照颜色分类水果和蔬菜的树中，表示苹果、西红柿和樱桃的那个叶都可以预计为“红色”，但由于可能出现绿苹果、黄西红柿和黑樱桃，所以会出现不同程度的置信度。

如果寄送一个新的目录，图6-1中的决策树把潜在的目录收件人分类为可能发来订单（1）或未必可能发来订单（0）。

图6-1中的树是利用SAS企业挖掘树查看器（SAS Enterprise Miner Tree Viewer）工具创建的。该图是依照数据挖掘中的惯例绘制——根在顶部、叶在底部，这也许暗示数据挖掘者应当更多地出去看看真树是如何生长的。每个结点在右上角标注有结点号，并在中间标注预测的类别；拆分每一结点的决策规则印在连接每一结点及其子结点的连线上；在根结点上按“生存期订单”拆分，左分支代表有6个或更少订单的顾客，右分支代表有7个或更多订单的顾客。

任何到达19、14、16、17或18叶结点的记录被分类为可能响应，因为在这种情况下预测类别为1。到达这些叶结点的路径描述了树中的规则，比如叶19的规则是，*如果该顾客已作了超过6.5次的订单且自上次订购以来时间少于765天，那么该顾客可能响应。*

细心的读者也许注意到：决策树中的一些拆分看起来没有差别，例如，结点17和18是按食品类别中包含物品所做出的订单数目区分的，但两个结点都标记为响应者。这是因为尽管在结点18中响应的概率要高于结点17，但它们都处于把记录分类为响应者的设定阈值之上。作为一个分类器，该模型只有两种输出，“1”和“0”。这种二元分类丢弃了某些有用的信息，而把我们引入了下一个主题：使用决策树来产生得分和概率。

6.1.2 评分

图6-2与图6-1的树相同，但使用了不同的树查看器，并且改进了设定值，该树现在带有了额外注解——即每一个结点在类1中记录的百分比。

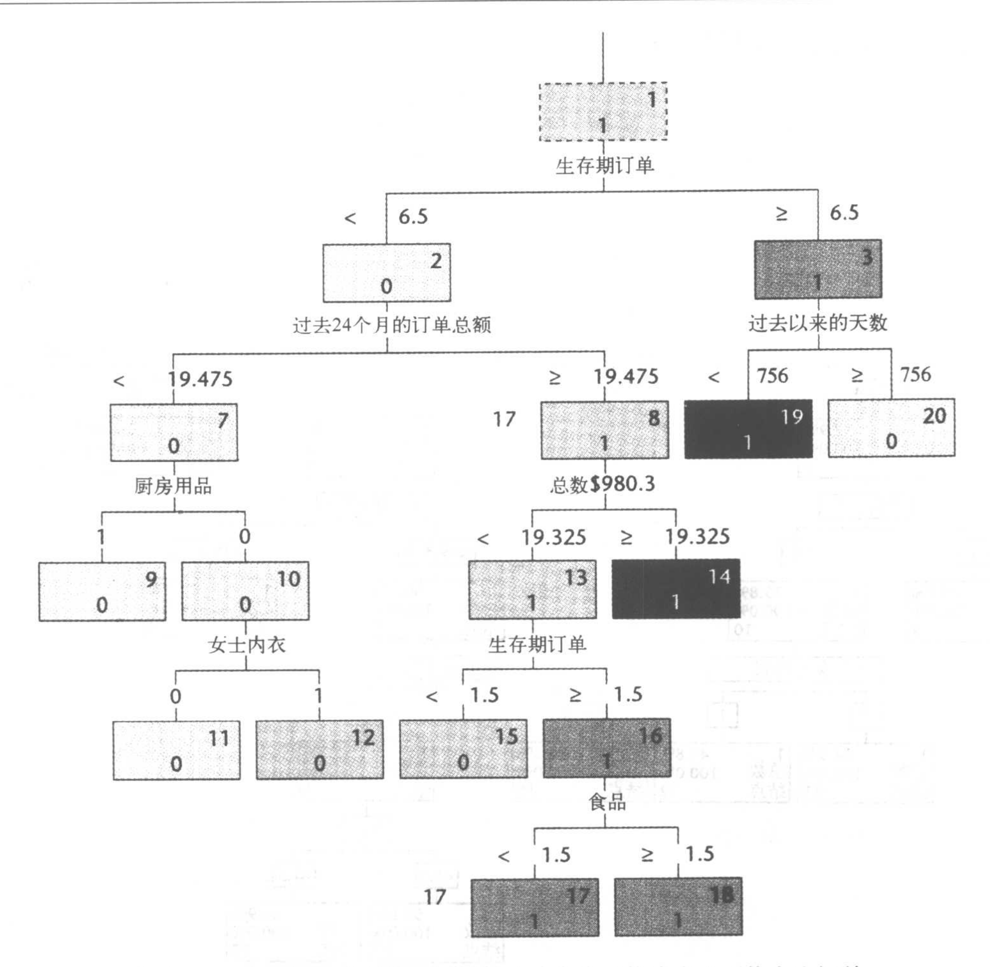


图 6-1 一个二元决策树把目录收件人分类为可能或未必可能发出订单

现在已经很清楚，该树描述了一个包含半数响应者和半数非响应者的数据集，因为根结点有 50％的比例。如同在第 3 章中描述的，这是一个典型的具有二元目标变量的响应模型训练集。具有超过 50％响应者的任何结点在图 6-1 中被标记为“1”，包括结点 17 和 18。图 6-2 阐明了这些结点之间的区别，在结点 17，52.8％的记录代表响应者，而在结点 18，这一比例是 66.9％。很明显，在结点 18 中的一条记录比在结点 17 中的一条记录更可能代表一个响应者。在期望的分类中，记录的比例可以当作一个得分使用，它常常比只进行分类更有用。对于二元结果，分类仅仅把记录拆分为两个组，而得分则可以对记录进行排序，从最可能到最不可能成为期望的分类成员。

对于许多应用而言，需要做的就是给出一个得分，按照得分排序列表，这足以选出用于投递的最佳 *N* 百分比，并且在列表的各种深度上计算提升度（lift）。然而对于另外的一些应用，比如要知道“A 比 B 是否更可能响应”，它就不够充分了，因为我们想知道来自 A 的响应实际可能性有多少。假定一个响应的先前概率是已知的，通过抽样数据建立的树结构产生的得分可以用于计算响应概率。换句话说，该模型能够适用于具有反映真实总体响应分布的预分类数据。这个被称为*逆向适应*（backfitting）的方法可以用来创建得分，即用该树叶结点处的分类比例来代表从相似总体中抽取的一条记录属于该类的概率。这些分类以及与此

相关的问题，在第 3 章中已作过详细讨论。

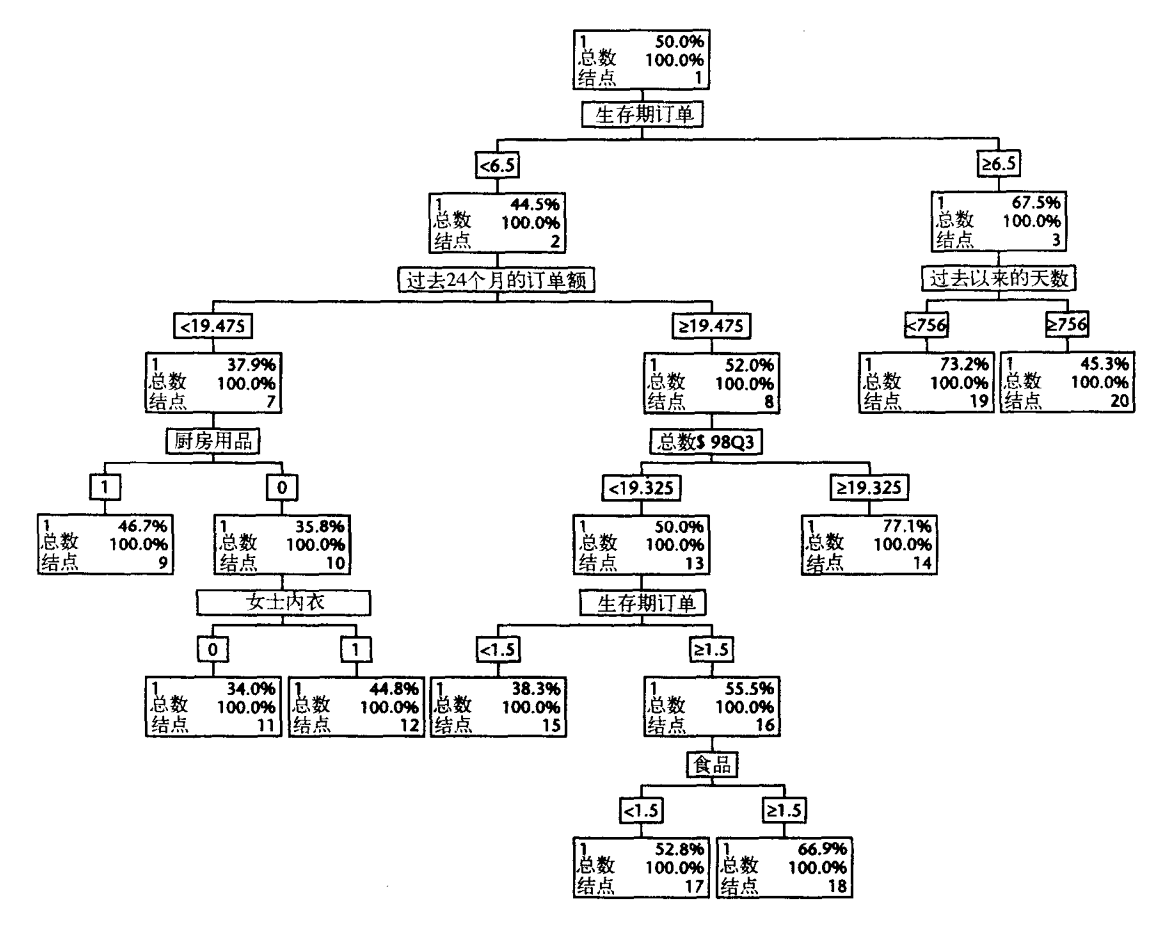


图 6-2 在决策树的每个结点标注记录在类 1 中的比例，显示该分类的概率

6.1.3 估计

假设重要的商业问题不是“*谁将响应*”，而是“*该顾客下一个订单的大小是多少*”，决策树同样能够回答这个问题。假定订单数量是预分类模型集的可用变量之一，在每一叶结点的平均订单大小可以作为满足该叶结点条件的任何未分类记录的估计订单大小。它甚至有可能使用数值型目标变量建立树，这样的树被称为*回归树*（regression tree）。被选中的树的每一次拆分，不是由于增加了分类变量的纯度，而是因为降低了每一子结点目标变量数值的方差（variance）。

事实上，树能够用于（有时确定就是）估计连续值，但这并不是一个好主意。决策树估算器能产生和树中叶子一样多的离散值。要估计连续变量，使用连续函数可能更好，回归模型和神经网络模型通常更适用于估计。

6.1.4 树以多种形态生长

在图 6-1 中的树是一个非均匀深度的二元树，就是说，每一个非叶结点有两个子结点，

并且叶结点与根结点距离并不都相同。在这种情况下，每一结点代表一个是或否的问题，其答案决定了一条记录向该树下一层进发的两条路径。因为任何多路拆分都能够表示为一连串的二元拆分，树实际上没必要有更多的分支数。不过，许多数据挖掘工具能够产生具有多于两个分支的树。例如，一些决策树算法依据分类变量拆分，对每个类生成一个分支，导致树在不同结点上有不同数目的分支。图 6-3 显示的分类问题与图 6-1 和图 6-2 相同，但它使用三路拆分和两路拆分得到树。

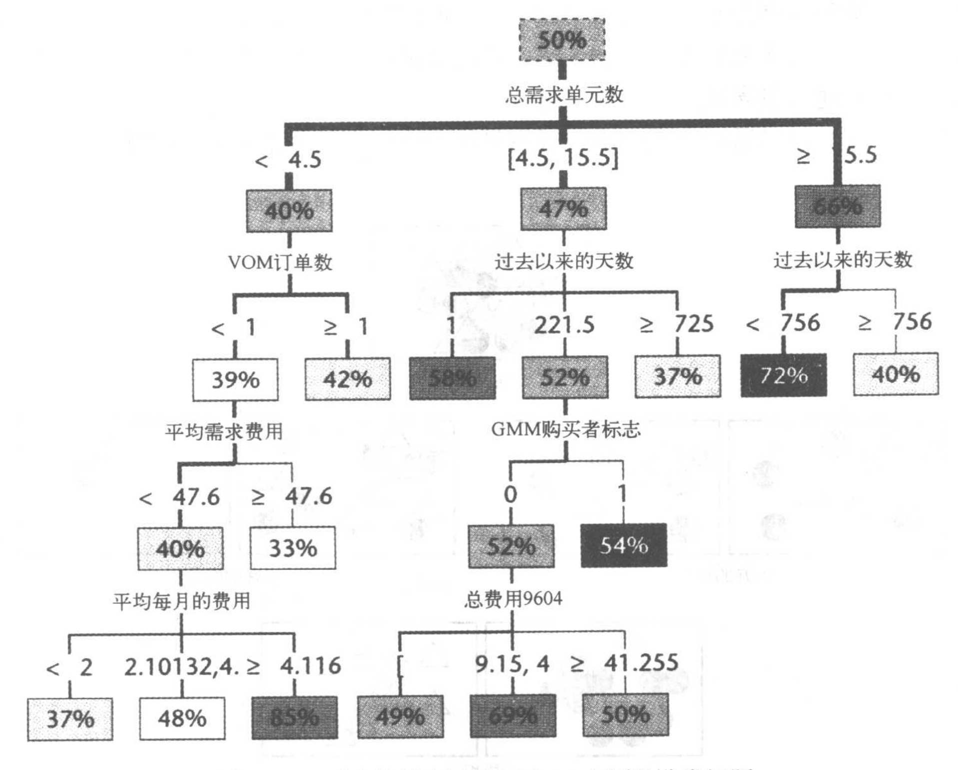


图 6-3 三元决策树应用于与图 6-1 相同的分类问题

提示： 在一个结点允许的分支数目和目标变量的类别数目之间没有关系。二元树（即每个结点有两路拆分）能够用于把记录分类为任何数目的类别，而多路拆分树能够用于分类二元目标变量。

6.2 决策树是如何长成的

尽管决策树核心算法有许多变体，但它们都具有相同的基本过程：相对于目标变量而言，每一新生结点比其原生结点有更高的纯度，通过这种方式，把数据重复地拆分为越来越小的群组。在本章的多数讨论中，我们假定变量是一个二元的分类目标变量，例如响应者和非响应者，在没有损失普遍适用性的前提下这样就简化了解释。

6.2.1 发现拆分

在这一过程的开始，有一个由预分类记录组成的训练集，换句话说，其中所有情形的目标变量值都是已知的。我们的目标是建立一棵树，基于输入变量的数值给新记录的目标字段

指派一个类（或归为每个类的可能性）。

通过在每一结点按照单一输入字段的功能拆分记录可以建立树，因而，首要的任务是确定哪一个输入字段会产生最佳拆分。最佳拆分可定义为：能够把记录很好地分割为不同的群组，使每个群组里的单个类成为主导。

这里用于评价可能拆分的度量是*纯度*（purity），下一节将更详细地讨论计算纯度的一些具体方法，不过，这些方法都试图达到相同的效果。对于所有这些方法，低纯度意味着该集合包含了各个类的典型分布（相对于父结点），而高纯度意味着单个类别的成员占主流。最佳拆分就是那些能够最大程度地增加该记录集纯度的拆分。好的拆分也会创建相似大小的结点，至少不要创建只包含很少记录的结点。

上述这些论点可以通过直观的方法很容易地看出。图 6-4 显示了一些好的拆分和不好的拆分。

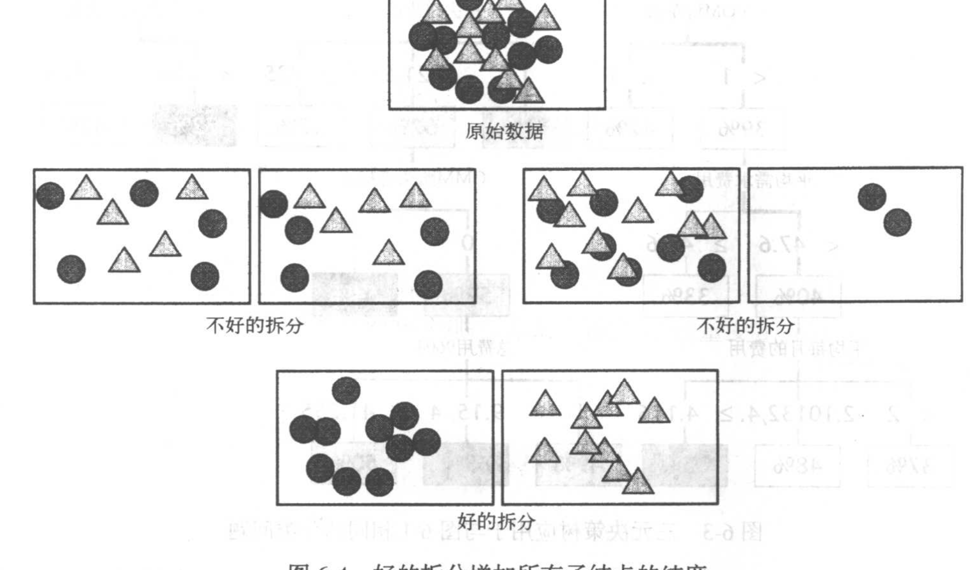


图 6-4　好的拆分增加所有子结点的纯度

第一个拆分不好，因为纯度没有增加。初始总体包含相等数量的两类点，在拆分后，每一个子结点仍然如此；第二个拆分也不好，因为尽管纯度稍有增加，纯结点只有很少的成员，并且较大结点的纯度只是比父结点稍微好一些；最后一个拆分是好的，因为它给出的子结点大小大体相同并且纯度比父结点高得多。

建树算法是一种穷举算法，方法是依次纳入每一个输入变量并测定由该变量建议的每一拆分所产生的纯度增加值，在尝试所有输入变量之后，产生最佳拆分的那一个被用于初始的拆分，生成两个或更多的子结点。如果不可能再分（因为只有太少的记录）或者没有能够改进拆分纯度，那么该算法结束于该结点，该结点变为一个叶结点；否则，该算法继续进行拆分，并在每一个子结点上重复进行，按这种方式重复自身的算法称为*递归算法*（recursive algorithm）。

按照目标变量对结点纯度的影响可以对拆分进行评价。这意味着选择一个适当的拆分标准依赖于目标变量的类型，而不是输入变量的类型。对于分类目标变量，无论输入变量生成

的拆分是数值型还是分类型（诸如 Gini、信息增益或卡方检验）都是适用的。类似地，对于一个连续的数值型变量，通过诸如方差归约或 F 测试等方法来评价该拆分是适用的，不管该输入变量产生的拆分是分类型的还是数值型的。

1. 按照数值型输入变量拆分

在对一个数值型输入变量寻找二元拆分方式时，变量在训练集中所取的每个值被当作一个拆分的候选值处理。对数值型变量的拆分采取“$X<N$”的形式，X（拆分变量）值小于某一常量 N 的所有记录被送往一个子结点，所有 X 值大于或等于 N 的记录被送往另一结点。在每次尝试性拆分之后，度量由该拆分引起的纯度的增加（如果有的话）。为提高效率，一些拆分算法工具实际上不是评估每个值，而是评估从这些数值中抽取的一个样本。

当给决策树评分的时候，数值型输入变量的惟一用途是把它们的数值和拆分点进行比较。它们从来不会像在许多其他类型的模型中那样，与权重相乘或求和到一起，这就导致了一个重要的结果，即决策树对数值变量的离群值（outlier）或倾斜分布（skewed distribution）不敏感，因为它只使用数值变量的等级，而不是它们的绝对数值。

2. 按照分类型输入变量拆分

拆分分类型输入变量的最简单算法是：对该分类变量所采用的每个类别创建新的分支。因此，如果颜色被选为拆分根结点的最好字段，且训练集包含有红、橙、黄、绿、蓝、靛、紫这些值的记录，那么在该树的下一层上将有 7 个结点。这种方法在某些软件包中已经被实际运用，但它常常产生很差的结果。高分支数会快速减少每个树的低层结点可用的训练记录总体，使得进一步拆分的可靠性降低。

一种更普遍的方法是把单独分类时预示相似结果的类组合到一起。更确切地讲，如果两个输入变量产生的输出变量分布差别不显著，那么这两个类别就可以合并。分布差别是否明显的常用测试方法是卡方检验。

3. 出现缺失值时的拆分方法

决策树最好的功能之一是它处理缺失值的能力。无论是输入字段数值型还是分类型，只要简单地将空值当作其自身分支上一个可能的值即可，这种方法比丢弃有缺失值的记录或试图归纳缺失值好得多。由于数值缺失而丢弃记录可能造成有偏离的训练集，因为包含缺失值的那些记录不可能是总体的随机样本。用归纳值替换缺失值则有这样的危险：有值缺失这一事实提供的重要信息在模型中将被忽略。我们已经看到许多案例，其中特定值为空的事实具有预言性价值。有一个这类案例，在追加的家庭层次人口统计学数据中，非空数值的计数与一个定期人寿保险服务的响应正相关。显然，与那些生活中留下更多空值字段的人相比，在 Acxiom 的家庭数据库中留下许多踪迹（通过买房子、结婚、登记产品和订阅杂志）的人们可能对人寿保险更感兴趣。

提示：决策树能够在有输入变量缺失值的情况下进行拆分。某值为空值的事实常常具有预言性价值，因此不要草率地筛选掉有缺失值的记录，或者试图将它们替换为归纳值。

尽管把空值作为一个单独的类别拆分常常很有价值，但很多数据挖掘产品还提供了其他替代方法。在 Enterprise Miner 中，每个结点存储几个可能的拆分规则，每一个都以一个不同的输入字段为基础。当产生最好拆分的字段中遇到空值时，该软件使用基于下一个最可用的输入变量的拆分作为替代。

6.2.2 生成完全树

首次拆分产生两个或者更多的子结点，然后以与根结点相同的方式继续拆分每一个子结点。所有输入字段又重新被看做候选拆分器，即使字段已经被用于拆分。但只呈现一个值的字段被排除在考虑之外，因为不可能使用它们建立拆分。已经在树的高处用作拆分器的分类字段可能相当快地变为单一值，这样对每一剩余字段的最好拆分就是确定的。当发现不再有拆分使给定结点的纯度显著增加时，或有结点中记录的数目达到某一预设的下界时，或者当树的深度达到某一预设的极限时，搜索那一分支的拆分就被放弃，该结点被标记为叶结点。

最终结果是，在树中任何地方都不可能发现更多的拆分，于是完全的决策树就生成了。就像我们将要看到的，这棵完全的树一般不能对新记录集最好地进行分类。

决策树建立算法通常始于在期望类别中试图发现能够最好地拆分数据的输入变量，在树的每一后继层，前一次拆分创建的子集本身按照最利于其工作的规则拆分，树继续生长，直到不可能发现更好的方法拆分新的记录。如果在输入变量和目标变量之间有十分确定的关系，这一递归拆分将最终产生一棵完全由纯叶结点组成的树。要造出这种例子很容易，但在市场交易活动或 CRM 应用中它们并不经常出现。

客户行为数据在输入和输出之间几乎从来不包含这样清晰的、确定性的关系。两个客户对于可用输入变量具有完全相同的描述，这个事实本身并不能保证他们将表现出相同的行为。一个关于邮购目录响应模型的决策树可能包含一个叶，代表年龄超过 50 岁、在去年之内购买了三次或者更多、生存期总花费超过 145 美元的女士。而到达该叶的典型客户是响应者和非响应者的混合体，如果在问题中该叶被标记为“响应者”，那么非响应者的比例就是这个叶的*误差率*，这个叶中响应者比例对总体响应者的比例之比是该叶的*提升度*。

发现确定性规则的一种可能情况是在数据中的模式（pattern）反映商业规则的时候，作者通过在 Caterpillar（一家内燃机制造商）的工作经历最终意会到了这一点。我们通过建立决策树模型来预测哪种保修索赔将被核准。那时候，该公司有一个政策，某些索赔是按照它自动支付的。结果很令人吃惊：在未使用过的测试数据上，该模型 100% 正确。换句话说，它发现了 Caterpillar 用于分类索赔的确切规则，而神经网络工具在这一问题上很少会成功。当然，发现已有的商业规则未必特别有用，然而它的确衬托出决策树在面向有规则问题时的有效性。

在许多领域，从遗传学到工业生产过程，确实存在潜规则，尽管这些潜规则可能很复杂，且会被嘈杂的数据淹没。当你怀疑存在潜规则时，决策树是一个很自然的选择。

6.2.3 度量决策树的有效性

决策树作为一个整体的有效性，可以通过把它应用于测试集（未用于建立该树的记录集合）观察其正确分类的百分比来确定。它提供了该树的总体分类误差率，但要注意该树单个分支的性质也很重要。穿过该树的每一条路径代表一条规则，其中的一些规则会比另一些好。

在每个结点处，无论叶结点还是枝结点，我们可以测量：

- 进入该结点的记录的数目
- 在每一类中记录的比例

- 如果这是一个叶结点，那么这些记录将被如何分类
- 在这一结点处记录正确分类的百分比
- 训练集和测试集分布之间的差异

其中特别令人感兴趣的是在该结点处记录被正确分类的百分比，令人惊讶的是，有时在树中高处的结点完成分类测试集的工作比底层的结点好。

6.3 选择最佳拆分的测试

有许多不同的方法可用于评估潜在的拆分。在机器学习领域开发的算法关注于拆分产生的纯度的增加，而那些在统计学领域开发的算法则关注于子结点分布上的统计学差异。改变拆分准则（splitting criteria）常常导致树的外观互不相同，但这些树却具有相似的性能，这是因为通常有许多性能非常相似的候选拆分。不同的纯度度量导致不同的候选者被选中，但既然所有的度量都试图捕捉同一个思想，最后得到的模型就会趋向于相似的性能。

6.3.1 纯度和发散性

本书第1版按照拆分引起的*发散性的降低*描述了拆分准则，在这一版中，我们改为按照*纯度的增加*来描述拆分准则，这好像更直观一些，这两个词所指的意思是相同的。纯度度量的范围可以从0（当样本中没有两项是在同一类别中）到1（当样本中所有项都在同一个类别中），用1减去纯度则转换为发散性度量。在评价决策树拆分时，有些度量方法习惯赋予纯的结点最低分数，有些则给纯的结点指派最高分数。本节把它们都用作纯度度量，目标是通过把被选度量最小化或最大化来优化纯度。

图6-5显示了一个好的拆分。父结点包含相等数目的亮点和暗点；左边的子结点包含9个亮点和1个暗点；右边的子结点包含9个暗点和1个亮点。毫无疑问，纯度增加了，但是这种增加该如何量化呢？这一拆分又如何与其他拆分相比较呢？这就需要一个纯度的正式定义，下面列出了其中的几个。

用于评价拆分分类目标变量的纯度度量包括：

- 基尼（Gini，也称总体发散性）
- 熵（entropy，也称信息增益）
- 信息增益比率
- 卡方检验

当目标变量为数值型时，一种途径是采用上述某一个方法，此外，还有两种方法都可用于数值型目标变量：

- 方差归约
- F测试

注意，选择适当的纯度度量方法取决于该*目标*变量是分类型还是数值型的，而输入变量的种类无关紧要，因为整个树是用相同的纯度度量方式建立的。在图6-5中演示的拆分可能通过一个数值型输入变量（年龄＞46）或通过一个分类型变量（STATE是CT、MA、ME、NH、RI、VT中的一个）来进行。不论拆分的类型如何，子结点的纯度都是相同的。

（译者注：美国的每个州通常可以用两个大写字母作为代码来表示，如CT——Connecticut，MA——Massachusetts。）

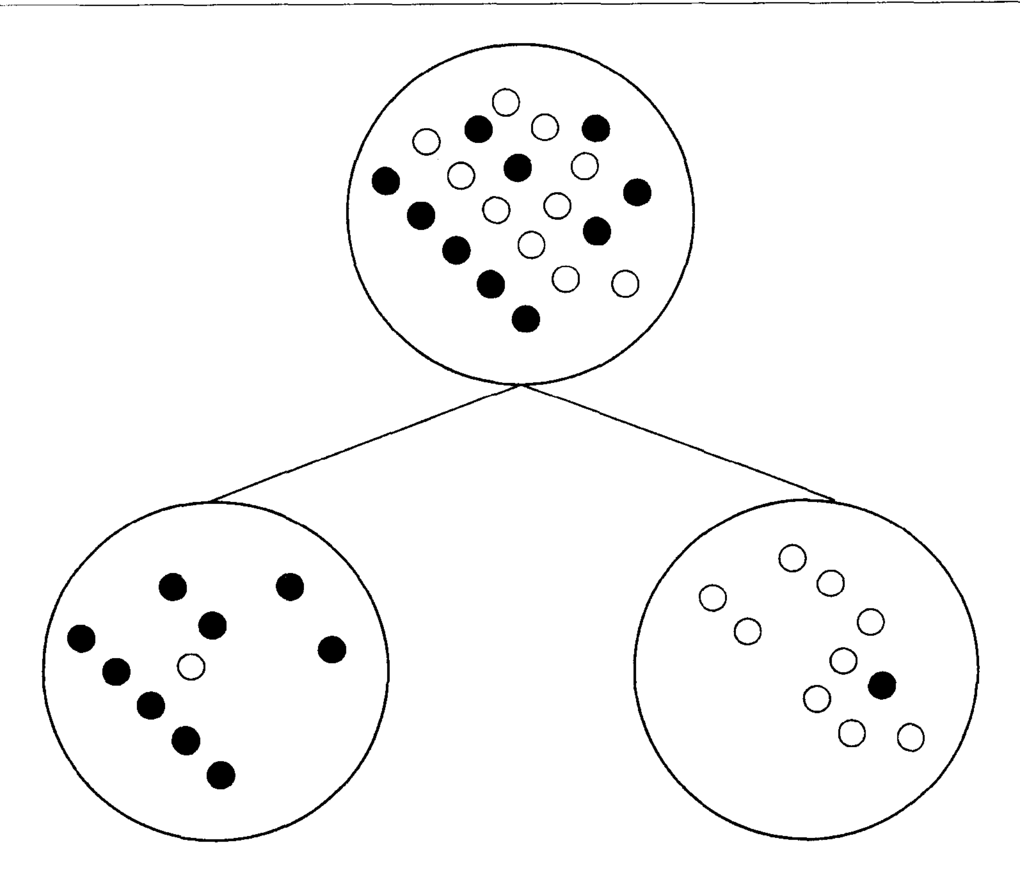


图 6-5 在一个二元分类变量上，好的拆分增加了纯度

6.3.2 基尼或总体发散性

一个通用的拆分标准被称为基尼，是以意大利统计学家和经济学家 Corrado Gini 的名字命名的，它也被生物学家和生态学家用于总体多样性研究，这种方法用于计算从相同的总体中随机选择的两项处于同一个类中的概率，对于一个纯的总体，此概率为 1。

结点的基尼值就是该类比例的平方之和。对于在图 6-5 中所示的拆分，父结点总体具有相等数目的亮点和暗点，可以预期的是，两个类中每个都具有相等数目的结点得分为$0.5^2+0.5^2=0.5$，因为通过随机选择和替换，两次选出同一个类的可能性是二中选一。所得到的每一个结点的基尼分数为 $0.1^2+0.9^2=0.82$。一个完美的纯结点基尼分数为 1，一个均等的平衡结点基尼分数为 0.5。有时该分数被加倍然后减 1，从而使它的值落在 0 和 1 之间。但在比较不同的分数以优化纯度时这样的变换没有差别。

要计算拆分的效果，可以把每一子结点的基尼分数乘以到达那个结点的记录的比例，然后把所有得到的数值求和。在这个例子中，因为记录是在拆分得到的两个结点之间被平均拆分的，并且每个结点都有相同的基尼分数，所以该拆分的得分与两个结点中任何一个的得分相同。

6.3.3 熵归约或信息增益

信息增益使用一个巧妙的想法来定义纯度。如果一个叶是完全纯的，那么在这个叶中的类很容易描述——它们都落入同一个类中。反之，如果一个叶是高度不纯的，那么描述它就复杂得多。信息论作为计算机科学的一部分，已设计出一个度量这种状况的方法，称为熵(entropy)。在信息论中，熵是对一个系统紊乱程度的度量。对信息论的全面介绍远远超出了本书的范围，对于我们的目的，直观的概念是描述特定的状况或结点需要的比特位数取决于可能结果集的大小。熵可被看做确定系统状态要进行的是或否问题的多少的一种度量方法，如果有 16 个可能的状态，它占用 $\log_2$ (16) 或者四个比特位来枚举它们或者识别其中的一个。附加的信息减少了确定该系统状态所需问题的数目，因此信息增益与熵归约意思是相同的。两个术语都可用于描述决策树算法。

对于某个指定决策树结点，熵是该结点所代表的全部类中，每个特定类的记录的比例乘以该比例以 2 为底的对数后的总和（实际上，这一总和通常乘以 -1 以便得到一个正数)。一个拆分的熵就是该拆分产生的所有结点的熵按照每个结点的记录所占比例的加权和。当熵归约被选作拆分准则时，算法搜寻的是能最大限度地减少熵（或者等价于信息增益）的拆分。

对于图 6-5 中所示的二元目标变量，单个结点熵的计算公式为：

$$-1*(P(\text{dark})\log_2 P(\text{dark})+P(\text{light})\log_2 P(\text{light}))$$

在这个例子中，P (dark) 和 P (light) 都是一半。把 0.5 代入熵的计算公式得：

$$-1*(0.5\log_2(0.5)+0.5\log_2(0.5))$$

第一项表示亮点 (light)，第二项表示暗点 (dark)，但因为亮点和暗点的数目相等，该式可简化为 $-1*\log_2(0.5)$ 也就是 +1。拆分产生的结点的熵是什么？其中一个结点有 1 个暗点和 9 个亮点，而另一个结点有 9 个暗点和 1 个亮点。显然，它们具有相同的熵，也即

$$-1*(0.1\log_2(0.1)+0.9\log_2(0.9))=0.33+0.14=0.47$$

为计算拆分后的系统熵的总和，用每个结点的熵乘以到达该结点的记录比例并把它们求和以得到平均值。在本例中，每个新结点接收了一半记录，因此总熵与每一结点的熵相同，即 0.47。因而，由于该拆分引起的总熵减少或信息增益为 0.53，这就是可用于比较这个拆分和其他候选拆分的数字。

6.3.4 信息增益比率

有一种拆分方法，为每个值创建一个单独分支来处理分类型输入变量，当熵拆分度量与这种方法结合时，可能会陷入麻烦。ID3 就属于这种情况，它是由澳大利亚研究者 J.Ross Quinlan 在 20 世纪 80 年代开发的一个决策树工具，已成为几种商业数据挖掘软件包的一部分。问题是，仅通过把大的数据分解到许多小的子集，出现在每个结点中类的数目趋向于下降，同时熵也会降低。完全归因于分支数的熵归约被称为拆分的本征信息 (intrisinc information)。(前面说过，熵被定义为每个分支的概率乘以该概率以 2 为底的对数，把所有分支之熵求和。）对于一个随机的 n 路拆分，每个分支的概率是 $1/n$，因此，单独归因于从一个

n 路拆分中得到的拆分的熵就是 $n*1/n\log(1/n)$ 或 $\log(1/n)$。正因为存在多路拆分的本征信息，如果对归因于拆分的本征信息没有任何校正，使用熵归约拆分准则建立的决策树会变得枝杈非常密集，而这样的具有许多多路拆分的密集树并不是我们想要得到的，因为这种拆分导致每个结点中有很少数目的记录，这是一个不稳定模型的构建办法。

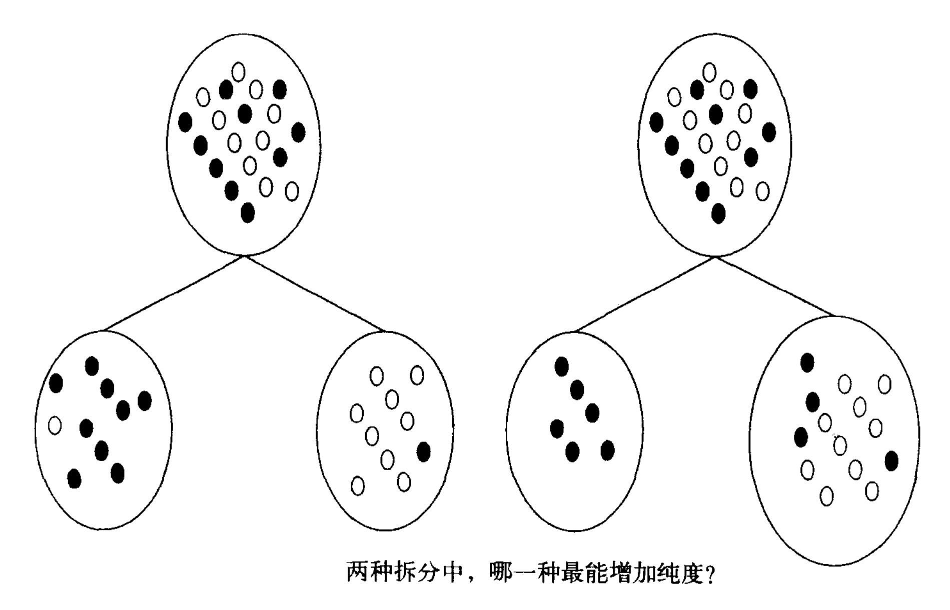
为解决这个问题，曾经使用信息增益的 C5 和其他 ID3 的后续版本现在使用一个比率，这个比率是由拆分引起的总的信息增益与可单独归因于评价拆分准则而创建的分支数的本征信息的比率。这种试验减少了在早期的决策树软件包中形成枝杈茂密树的趋向。

6.3.5 卡方检验

像在第 5 章中描述的那样，卡方（X^2）检验是由英国统计学家 Karl Pearson 在 1900 年提出的用于测试统计学显著性的方法。卡方被定义为在多重不相交样本中，某个事件的期望频率和观察频率的标准差的平方和。换句话说，该测试所度量的是在样本间观察到的差异只归因于偶然性的概率。当用于测量决策树拆分纯度时，较高的卡方数值意味着差异更显著，不能仅仅归因于偶然。

使用基尼和熵比较两个拆分

考虑下列两个拆分，如下图所示。在两个拆分中，黑点和亮点总体开始时非常平衡，每种都有 10 个。一种拆分产生与图 6-5 相同的两个大小相等的结点，其中一个包含 90% 的暗点，而另一个包含 90% 的亮点；第二种拆分产生一个含有 100% 的纯暗点的结点，但只有 6 个点；而另一个包含 14 个点，但其中只有 71.4% 是亮点。



两种拆分中，哪一种最能增加纯度？

使用基尼评价这两个拆分

像正文中阐释的那样，在第一种拆分中的两个子结点，每一个的基尼得分是 $0.1^2+0.9^2=0.820$。因为子结点大小相同，这也就是该拆分的得分。

第二种拆分又如何？左边结点的基尼得分是 1，因为只出现了一个类别，右边结点的基尼得分是：

$Gini_{right} = (4/14)^2 + (10/14)^2 = 0.082 + 0.510 = 0.592$

这一拆分的基尼得分是:

$(6/20)\ Gini_{left} + (14/20)\ Gini_{right} = 0.3 * 1 + 0.7 * 0.592 = 0.714$

既然第一个拆分的基尼得分(0.820)比第二个拆分的基尼得分(0.714)大,使用基尼准则建立的树将更倾向于产生两个几乎很纯的子结点的拆分,而不是产生一个完全纯的子结点加上一个更大但不很纯的子结点拆分。

使用熵评价这两个拆分

像在正文中计算的那样,父结点的熵是1。第一个拆分的熵也在正文中计算过,是0.47,因此第一种拆分的信息增益是0.53。

第二种拆分的信息增益是多少?左边的子结点是纯的,因此熵为0。对于右边的子结点,熵的计算公式是

$-(P(dark)\log_2 P(dark) + P(light)\log_2 P(light))$

因此右边子结点的熵是:

$Entropy_{right} = -((4/14)\log_2(4/14) + (10/14)\log_2(10/14)) = 0.516 + 0.347 = 0.863$

这种拆分的熵是所生成结点熵的加权平均值。在本例中为:

$0.3 * Entropy_{left} + 0.7 * Entropy_{right} = 0.3 * 0 + 0.7 * 0.863 = 0.604$

从父结点的熵(为1)中减去0.604得到0.396的信息增益,这比第一种拆分的信息增益0.53少,所以在这个案例中,熵拆分准则也更倾向于第一种拆分,而不是第二种。与基尼得分相比较,熵准则对更纯的结点确实有更强的倾向,即使该结点很小。在确实有清晰的潜在规则的领域,这也许是适合的,但在诸如对市场营销服务做出响应这种领域,它会导致不太稳定的树。

例如,假定目标变量表示的是在产品引导期客户是否将续约的二元标记,拆分是按"获取渠道"建立的,这是一个有直接邮寄、长途电话和电子邮件三个类别的分类变量。如果获取渠道对更新率没有影响,我们可以预期每个类的更新数目与通过该渠道获取的顾客数目成正比。对于每个渠道,卡方检验分值可以如下计算:从实际观察的更新中减去期望的更新数目,计算二者差的平方,并除以期望值数目,最后把每个类的对应数值加在一起就得到了该分值。正如第5章中描述的那样,卡方分布提供了一种把卡方检验分值转化为概率的方法。在决策树中测量拆分的纯度,使用这个分数就足够了,高得分表示该拆分可以成功地把总体拆分为有显著分布差异的次级分组。

卡方检验把它命名为CHAID,是由John A. Hartigan在1975年首次发表的有名的决策树算法,这个缩写词代表卡方自动交互检测器(Chi-square Automatic Interaction Detector)。顾名思议,CHAID的最初动机就是为了检测变量之间的统计学关系,它通过建立决策树做到这一点,因此这个方法也已经被用作分类工具。CHAID使用卡方检验有以下几种方式——首先是合并目标变量上没有重要差异的类,然后选择最佳拆分,最后确定在一个结点上是否有必要执行任何另外的拆分。在研究领域,目前流行的是尽可能少用续拆分的方法(仅仅当可能有用的时候才使用)而倾向于包含修剪的方法。但有一些研究者仍然喜欢原始

的 CHAID 方法而不喜欢修剪。

卡方检验应用于分类变量，所以在经典 CHAID 算法中，输入变量必须是分类型。连续变量必须归档或替换为顺序的类别，例如高、中、低。当前的一些决策树工具，例如 SAS Enterprise Miner，使用卡方检验进行分类变量的拆分，而使用另一项统计学测试（即 F 测试）对连续变量进行拆分。同样，即使当拆分不具有统计学方面的显著差别时，一些 CHAID 工具仍继续建树，然后应用修剪算法把树裁剪回来。

6.3.6 方差归约

前面四个纯度度量方法都是应用于分类型目标的。当目标变量是数值型时，一个好的拆分应当减少目标变量的方差。前已述及，方差度量的是总体中接近于均值的趋向。在具有低方差的样本中，大多数数值非常接近均值；在具有高方差的样本中，许多数值远离均值。方差的实际公式是标准差平方和的均值。尽管方差归约的拆分准则是针对数值变量的，但它仍然能够用于图 6-5 中的暗点和亮点，可以通过把暗点视为 1、亮点视为 0 来应用。父结点的均值很明显是 0.5，20 个观察值的每一个都与均值有 0.5 的差异，因此方差为 $(20*0.5^2)/20=0.25$。拆分以后，左边的子结点有 9 个暗点和一个亮点，因此该结点的均值是 0.9。9 个观察值与均值有 0.1 的差异，1 个观察值与均值有 0.9 的差异，因此方差为 $(0.9^2+9*0.1^2)/10=0.09$。因为拆分后的两个结点方差都是 0.09，拆分后总的方差也是 0.09，由于拆分引起的方差归约是 $0.25-0.09=0.16$。

6.3.7 F 测试

另一个可用于数值目标变量的拆分准则为 F 测试，它是以另一位著名的英国统计学家、天文学家和遗传学家 Ronald.A.Fisher 的名字命名。尽管（也可能是由于）Fisher 和 Pearson 兴趣范围有很大部分的重叠，据说他们并不来往，但 Fisher 针对连续变量的测试所做的工作就是 Pearson 针对分类变量的卡方检验所做的工作，它提供了一个度量概率的方法，可以用于度量具有不同均值和方差的样本实际取自同一个总体的概率。

在样本的方差和被取样总体的方差之间有一个很好理解的关系（事实上，只要样本大小合理并且是从总体中随机抽取的，从样本方差可以很好地估计总体方差。很小的样本——少于 30 条观测值——通常比它们对应的总体具有更高的方差）。F 测试观察的是两个总体方差估计之间的关系——一个是抽出所有的样本并计算组合样本的方差，另一个是计算中间样本方差作为样本平均方差。如果诸多不同的样本是从同一个总体中随机抽取的，这两个估计值应当非常接近。

F 分值是两个估计值之比，把中间样本估计值除以抽取的样本估计值即可求出该分值。该分值越大，样本就越不可能全部是从同一个总体中随机抽取的。在决策树环境下，一个大的 F 分值表明拆分已成功地把总体拆分为具有显著分布差异的分组。

6.4 修剪

如前所述，只要能找到新的拆分，能够改善把训练集中的记录分割为更纯的子集的能力，决策树就会继续长高。这样的决策树已针对训练集进行优化，因而去掉任何叶子都会增加该树在训练集上的误差率。这是否暗示着完整的树将完成最好的新数据集的分类工作呢？

当然不是！

决策树算法首先在有大量记录的根结点处做出最好的拆分，随着结点变得越来越小，一个结点上特定训练记录的特性开始支配该过程。理解这一点的一种方式是该树在大结点发现通用模式，在小结点发现训练集的具体模式。换句话说，该树过度适应（overfit）于该训练集，结果将是一个不会做出好的预测的不稳定的树。解决这个问题的对策是通过一个称为修剪的过程，它合并小的叶子从而排除不稳定的拆分。下面详细讨论三种通用的修剪算法。

6.4.1 CART 修剪算法

CART 是一个流行的决策树算法，由 Leo Breiman、Jerome Friedman、Richard Olshen 和 Charles Stone 在 1984 年首先发布，CART 是英文词组"Classification and Regression Trees"（分类与回归树）的首字母缩写词。CART 算法产生二元树，而且只要发现新的拆分能增加纯度就继续拆分。如图 6-6 所示，在一个复杂的树内部，有许多较简单的子树，每一个子树代表在模型复杂性和训练集误分类率之间的一种折衷。CART 算法把这样一些子树的集合视为候选模型，这些候选子树被应用于验证集，具有最低验证集误分类率的树被选作最终模型。

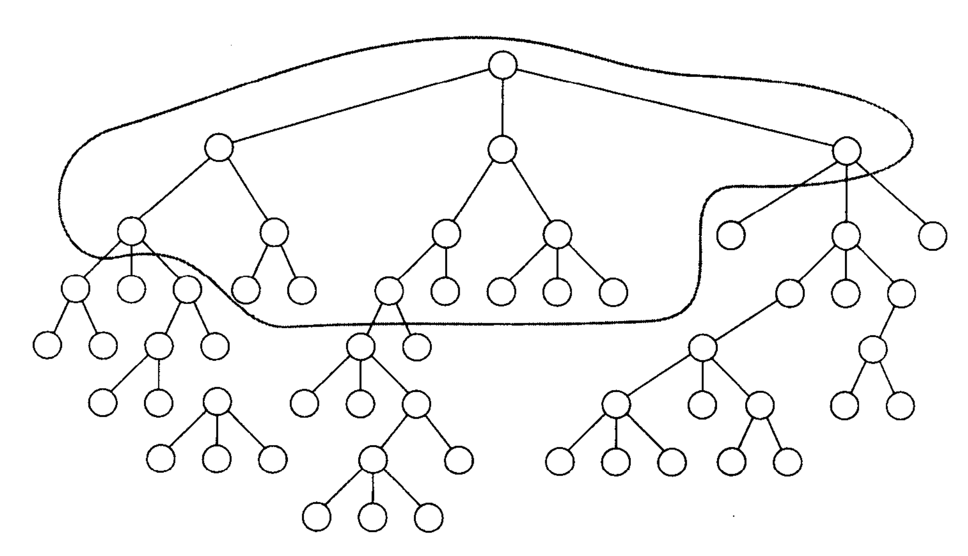


图 6-6 在复杂树的内部，有更简单、更稳定的树

1. 创建候选子树

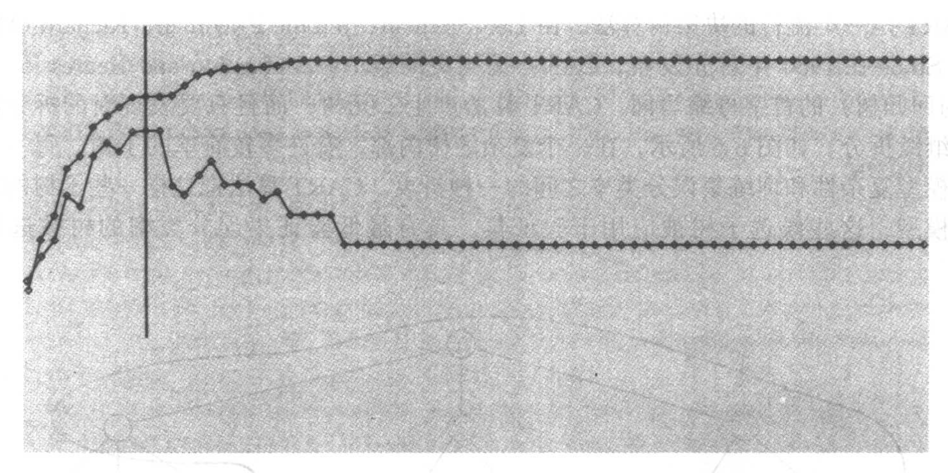
CART 算法通过重复修剪过程识别出候选子树，目标是首先修剪那些每个叶提供最少附加预言性能力的分支。为了识别这些最没用的分支，CART 依靠一个称为*调整误差率*（adjusted error rate）的概念。这种方法基于该树中叶的数目给出复杂性罚分，从而在训练集上增加每个结点的误分类率。调整误差率可用于识别弱的分支（误分类率不够低，因而不能超过罚分的分支），并做上修剪标记。

在训练集和验证集上比较误分类率

验证集（validation set）上的误差率应当比训练集上的误差率大，因为训练集是用于建立模型中的规则的。然而，在误分类误差率中的一个巨大差值是一个不稳定模型的征兆。这

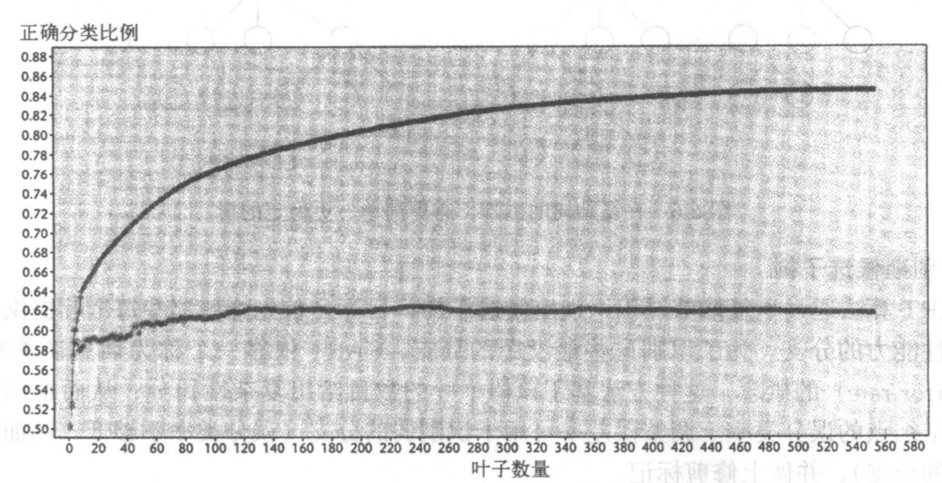
一差值能够用 SAS Enterprise Miner 产生的如下三个图所示的几种方式表示出来。该图给出的是在决策树中候选模型正确分类的记录的百分比，有较少结点的候选子树在左边，有更多结点的在右边。这些图显示了正确分类的百分比而不是误差率，因此它们与本书中其他地方所示的类似图表的方式是相互颠倒的。

像预期的那样，第一个图表表明在训练集上，当树的结点越来越多时候选树表现得越来越好，当表现不再改善时，训练过程停止。然而在验证集上，候选树达到一个峰值，然后随着树变大性能开始下降。最优树是在验证集上起作用的那棵树，挑选很容易，因为峰值轮廓分明。



这个图表显示出在验证集正确分类的百分比图中的一个清晰的拐点

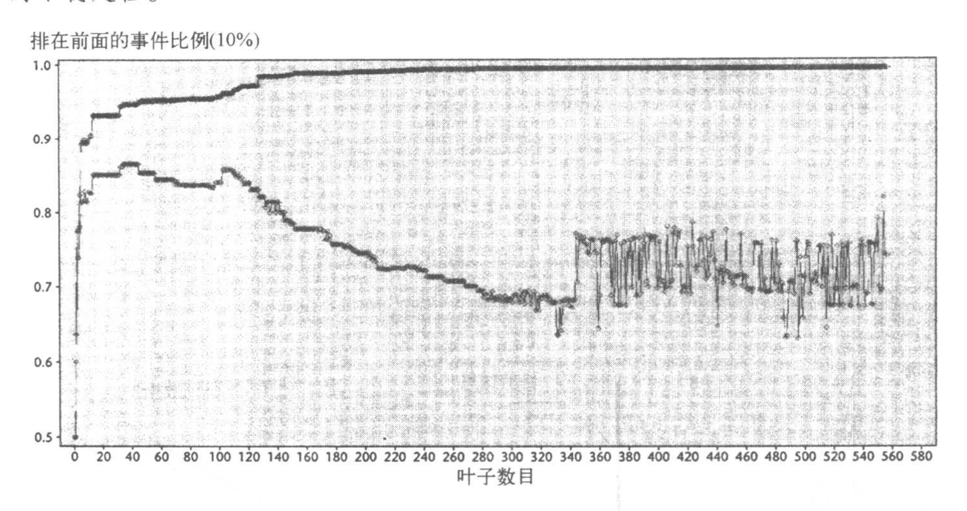
但有时并没有清晰的分界点。换句话说，当树变大时，候选模型在验证集上的表现从来没有完全到达最高值。在这种情况下，修剪算法选择整个树（可能的最大子树），如下图所示。



在这个图表中，验证集中正确分类的百分比早早地变平，一直维持在远低于训练集中正确分类的百分比处

最后的例子可能是最值得关注的，因为当候选树变大后，验证集上的结果变得不稳定，

不稳定的原因是叶子太小了。在这棵树中有一个叶子，它包含训练集中的三条记录，并且所有三条记录的目标值都为1——一个完美叶子的例子。然而，在验证集中，落到那里的一条记录值为0，该叶是100%错误的。当树生长得更复杂，更多的这类小叶子被包括进来，导致如下的不稳定性。



在这个图表中，随树的复杂度增加，验证集上的正确分类百分比降低，并最终变得混乱

最后两个图是不稳定模型的例子。避免这类不稳定性的最简单方法是确保叶子不要变得太小。

调整误差率公式是：

$$AE(T)=E(T)+\alpha leaf_count(T)$$

其中α是调整因子系数，随新子树创建逐步增大。当α为零时，调整误差率等于误差率。为找到第一棵子树，随α逐步增加，包括根结点在内的所有可能的子树的调整误差率都被做出评价。当一些子树的调整误差率变得小于或等于整个树的调整误差率时，我们就找到了第一棵候选子树，即α_1；所有不是α_1的组成部分的分支被剪除，然后过程重新开始，α_1树被修剪创建一个α_2树；当树被一路修剪至根结点时该过程结束。每一个产生的子树（有时被称为*alphas*）是最终模型的候选者。注意：所有的候选者包括根结点，最大的候选者就是整个树。

2. 挑选最佳子树

接下来的任务是从候选的众多子树中选择在新的数据上工作最好的子树。当然，这是验证集的用途之所在，每一个候选子树被用于分类验证集中的记录，执行这一任务时给出最低的总误差率的那棵树被宣布为获胜者。获胜的子树已经对消除训练过度的效果做了足够充分的修剪，但又不怎么损失有价值的信息。图6-7中的图形演示了在分类准确性上修剪的效果。技术旁白更详细深入地讨论了这一点。

因为这一修剪算法是完全基于误分类率的，没有考虑每一种类别的概率，所以它把所有叶子给出相同分类的任何子树替换为做出同样分类的公共父结点。在目标是选出一个很小比例记录（例如最高的1%或10%）的应用中，这一修剪算法可能损害树的性能，因为一些被

删除的叶子可能包含很高比例的目标类别。某些工具，例如 SAS Enterprise Miner，允许用户针对这种情况对树做出最优的修剪。

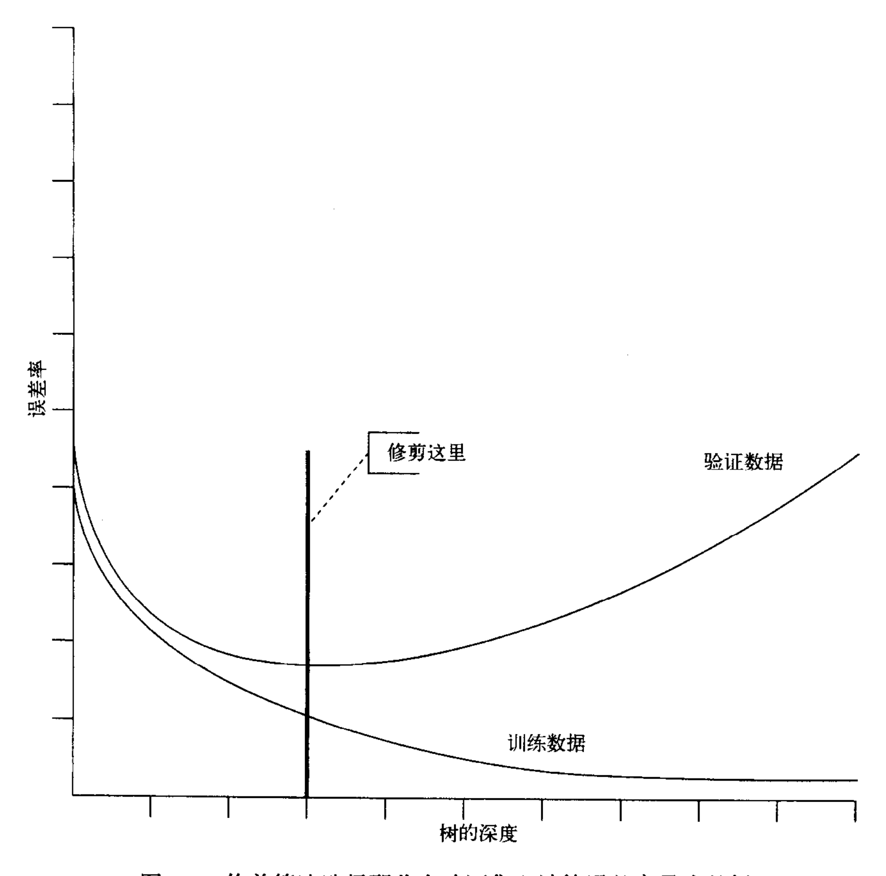


图 6-7 修剪算法选择那些在验证集上计算误差率最小的树

3. 使用测试集评价最终的树

在验证集中的记录被应用于分类任务时，获胜的子树是以其总误差率为基础被选出的。但是，当我们期待选中的子树在用于其他数据集时仍将表现最好，使它被选中的误差率可能有一点夸大它的有效性。可能有很大数量的子树表现得与选中的树一样好，在某种程度上，这些子树中在验证集上给出最低误差率的那棵树也许是从这些特定的记录集合中“凭侥幸”被选中的。正因为如此，像在第 3 章中解释的那样，被选中的子树被应用于与验证集和训练集都不相交的第三个预分类数据集，这个数据集被称为测试集。在测试集上得到的误差率被用于预测将选中树所代表的分类规则应用到未分类数据时的期望性能。

警告：不要利用在验证集上的提升度或误差率评价一个模型的性能。与训练集一样，它已参与了模型创建，因此将夸大模型的准确度。要始终在那些与训练集和验证集出自同一总体但没有以任何方式用于创建模型的测试集上测量模型的准确度。

6.4.2 C5 修剪算法

C5 是澳大利亚研究者 J. Ross Quinlan 演化和精修多年的决策树算法的最新版本。它的

一个早期的版本 ID3，是在 1986 年发布的，在机器学习领域非常有影响，它的后续版本被应用于几种商业数据挖掘产品中（ID3 这个名字代表“Iterative Dichotomiser 3”（迭代二分器 3），我们还没有听到对 C5 这个名字的解释，但是我们能够猜出 Quinlan 教授的教育背景是数学而不是市场营销）。作为商业产品，C5 可以从 RuleQuest（www.rulequest.com）购买。

用 C5 生成的树与那些用 CART 生成的树很相似（尽管和 CART 不同，C5 是在分类变量上进行多路拆分）。类似于 CART，C5 算法首先生成一棵过度适应的树然后把它修剪回来创建一个更稳定的模型，但修剪的策略却非常不同。C5 并不使用验证集从候选子树中做出选择，因为用于生成该树的相同数据也用于判定该树应当如何修剪。这或许反映出该算法起源于学术界，在过去，大学研究者很难有时间把精力放在用作训练集的大量真实数据上。因此，他们花更多的时间和精力，试图从穷尽的数据集中尽可能发掘到哪怕是最后的一点信息——在商业界中数据挖掘者是不会遇到这个问题的。

保守式修剪

C5 通过检查每一结点的误差率修剪树，并假定真实的误差率实际上已经足够差。如果 N 个记录到达一个结点，其中 E 个是分类错误的，那么该结点的误差率为 E/N。现在生成树算法的全部要点是最小化这一误差率，因此算法假定在所能够做到的范围内，给出最小 E/N 的树是最好的。

C5 使用带有统计学抽样的类推法给出在一个叶上可能出现的最坏误差率的评估值。该类推法通过如下方式工作：把该叶上的数据视为表示一系列尝试的结果，其中的每个尝试能够有两个可能的结果（首或尾是常见的结果）。正如已经发生的那样，至少从 1713 年起，即 Jacques Bernoulli 的著名二项式定理发布的那年，统计学家就已经在研究这种特殊的情形。因此有现成的公式可用于确定在 N 次尝试中观察到 E 次出现有多大意义。

特别是，有一个针对某个给定置信水平的公式，可以给出置信区间——E 的预期数值的范围。C5 假定在训练集上观察的误差数目是该范围的低端，并代入高端来得到一个叶的推算误差率，即在未见数据上的 E/N。结点越小，误差率越高。当一个结点的高端误差估计值小于其子结点的误差估计值时，该子结点被修剪掉。

6.4.3 基于稳定性的修剪

CART 和 C5（实际上也包括作者用过的所有商业决策树工具）使用的修剪算法有一个问题，它们未能修剪一些明显不稳定的结点。在图 6-8 中高亮显示的拆分是一个很好的例子。该图是用 SAS Enterprise Miner 观察一棵树时的默认设置生成的，每一结点的左边数目显示了在训练集上发生的情况，结点右边的数目显示了在验证集上发生的情况，这一特定的树试图识别流失者。当只考虑训练集数据时，高亮显示的分支看来工作良好，流失者的集中度从 58.0%升至 70.9%。不幸的是，当完全相同的规则应用于验证集时，流失者的集中度实际从 56.6%下降至 52%。

一个模型的主要目的之一就是在先前未见的记录上做出一致的预测，不能达到这个目标的任何规则都应当从模型中去除。许多数据挖掘工具允许用户手动修剪决策树。这是一个有用的功能，但作为一个选择，我们期待能够出现自动的基于稳定性进行修剪的数据挖掘软件。这一软件需要对“验证集结果的分布看起来不同于训练集结果的分布”这样问题的拒绝

拆分有更少的主观判别。一种可能性是使用统计学显著性测试，例如卡方检验或者比例差值，当置信水平低于某一用户定义的限度时，拆分将被修剪，于是只有那些在验证集上有一定置信度（比如说 99%）的拆分会保留下来。

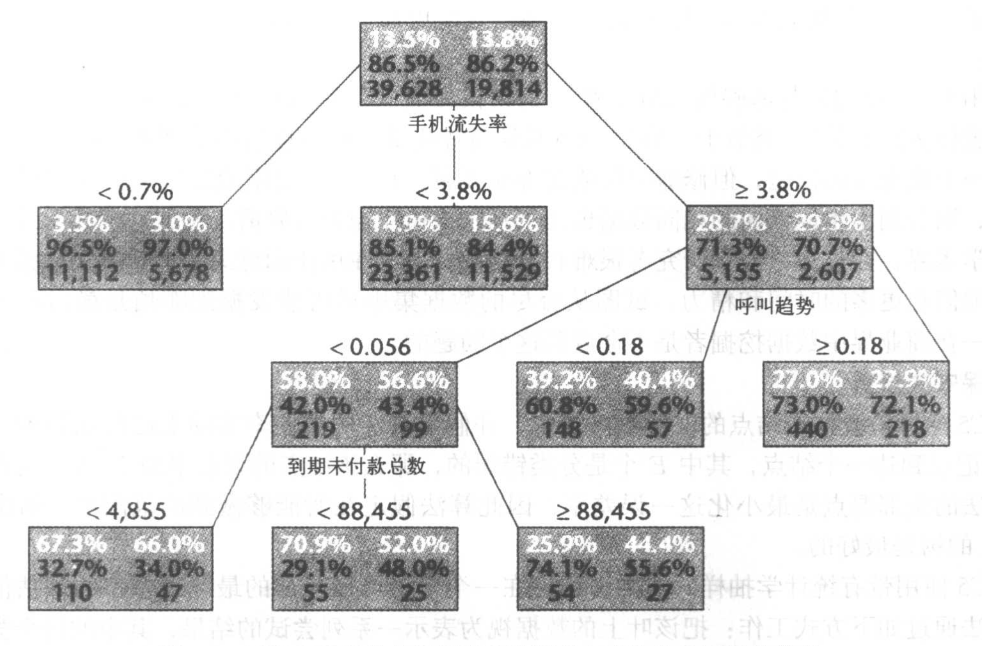


图 6-8 不稳定的拆分在训练集和验证集上产生非常不同的分布

警告：小结点导致大问题。不稳定决策树模型的一个共同原因是允许结点有太少的记录，大多数决策树工具允许用户设定一个最小结点大小。作为一个经验规则，那些大约少于 100 条训练集记录的结点可能是不稳定的。

6.5 从树中提炼规则

当决策树主要用于产生得分时，很容易忘记决策树实际上是一系列规则的集合。如果数据挖掘工作的目的之一是获得对问题领域的了解，这种得分对在决策树中把非常混乱的规则简化为小的更可理解的集合是有用的。

在期望输出是一系列规则集合时还有其他的情形。在 *Mastering Data Mining* 一书中，我们描述了决策树对一项工艺流程性能改善问题的应用，也就是防止某一类型的印刷缺陷。在那个案例中，数据挖掘工程的最终产出是有数条简单规则的小集合，这些简单的规则可以张贴在印刷车间的墙上。

当决策树用于产生得分时，拥有大量数目的叶是有利的，因为每个叶会产生一个不同的得分。当目标是产生规则时，规则越少越好。幸运的是，通常可以把复杂的树分解为小的规则集。

在这个方向上的第一步是组合那些造成相同分类的叶的路径。在图 6-9 中的部分决策树产生如下规则：

- 观看比赛且主队获胜，并与朋友外出，则喝**啤酒**。
- 观看比赛且主队获胜，并坐在家里，则喝**汽水**。

- 观看比赛且主队失败，并与朋友外出，则喝**啤酒**。
- 观看比赛且主队失败，并坐在家里，则喝**牛奶**。

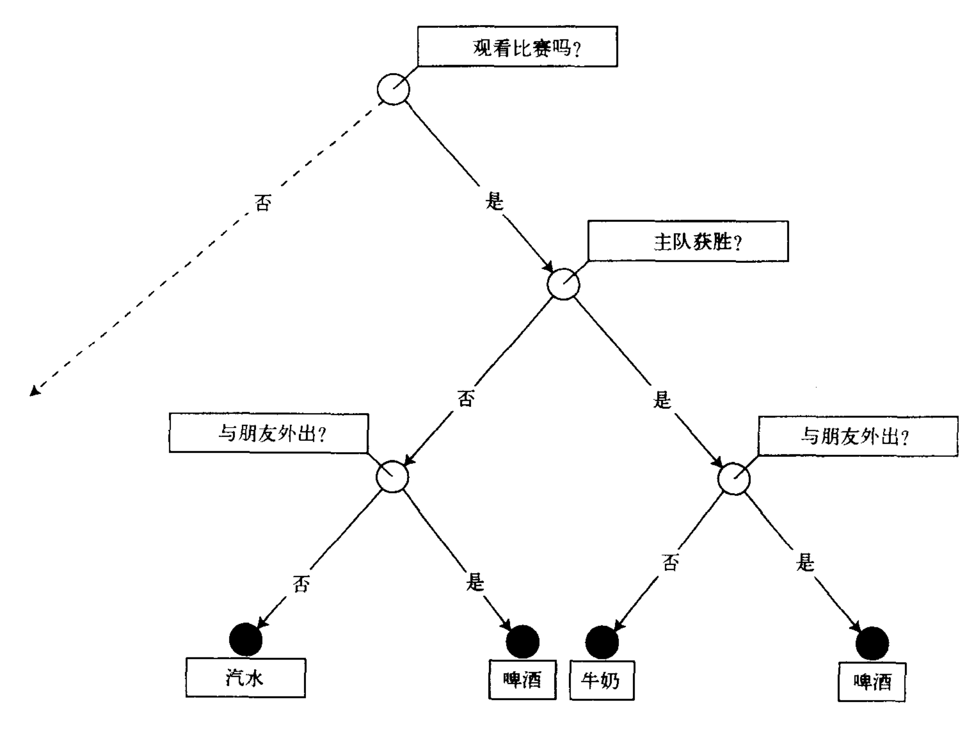


图 6-9 多重路径导致相同的结论

预言喝啤酒的两个规则能够通过删除主队是胜或败的测试而组合起来。该测试对于区分喝牛奶和喝汽水时是重要的，但并不带有与啤酒消费相关的信息。新的、简单一些的规则是：

观看比赛并与朋友外出，则喝**啤酒**。

迄今为止，没有争议是因为没有丢失信息，但 C5 的规则发生器在走向深入。它企图通过删除子句概化每个规则，然后，使用早先在用于修剪该树时相同的悲观误差率假设，把新的简洁规则的预计误差率和原始的相比较。常常，几个不同叶的规则概化为相同的规则，于是这一过程比有叶的决策树给出更少的规则。

在决策树中，每一条记录恰恰终结于一个叶，因此每条记录有一个确定的类。然而在规则概化过程之后，或许有并非互不相交的规则和不被任何规则覆盖的记录。当有多于一条规则适用时，简单地挑选出一条规则能够解决第一个问题，第二个问题需要引入一个分配给没有被任何规则覆盖的任何记录的*默认类别*，典型地，最经常出现的类别被选作默认类别。

一旦建立起概化规则的集合，Quinlan 的 C5 算法把每个类的规则组合在一起，并把对规则集合作为一个整体的准确度似乎贡献不大的那些规则排除掉，最终结果是获得小数目的容易理解的规则。

6.6 考虑成本

在迄今为止的讨论中，误差率是评价规则和子树相互匹配的惟一度量。然而在许多应用中，误分类的成本在类与类之间经常大不相同。毫无疑问，在一次医疗诊断中，假阴性比假

阳性可能更有害。通过深入调查，一个令人恐慌的巴氏早期癌变探查涂片结果被证明是假阳性的，远比癌症未被发现要好。我们可以用误分类概率乘以权重的成本函数表示误分类的成本，有几种工具允许利用这样的成本函数而不是误差函数建立决策树。

6.7 决策树方法的进一步修正

尽管在多数的商业数据挖掘软件包中找不到，还是有一些对基本决策树方法的重要的修正值得讨论。

6.7.1 每次使用多于一个字段

多数决策树算法测试单一变量来实施每次拆分。这种方法可能因为几个原因而出现问题，至少会导致形成结点数目过多的树。多余的结点令人关注，是因为只有到达给定结点的训练记录可用于推测其下的子树，每个结点包含的训练样本越少，最后得到的模型越不稳定。

假定我们对年龄和性别两者都是重要指示器的条件感兴趣，如果根结点是按年龄拆分的，那么每个子结点只包含大约一半的女士。如果初始的拆分是按性别的，那么每个子结点只包含大约一半的老人。

现在已开发出若干算法，允许多重属性组合以形成拆分器。有一种技术可以形成特征的布尔逻辑乘积以降低树的复杂性。在发现形成最好拆分的特征之后，该算法就寻找与最初选出的特征组合能够最大程度地改善拆分的那一特征。只要在结果拆分中继续具有统计学意义的显著改善，特征就会继续增加。

这一过程能导致形成一个更有效的分类规则。举例来说，假设我们，按照投票活动是否获得全体通过来分类投票结果。为了简单起见，考虑只有三个投票者的情况（简化程度只是增加投票者的数目）。

表 6-1 包含三个投票的所有可能组合，增加一列指示结果的全体一致性。

表 6-1 三个投票者的所有可能的投票组合

第一个投票者	第二个投票者	第三个投票者	全体一致吗?
Nay	Nay	Nay	真
Nay	Nay	Aye	假
Nay	Aye	Nay	假
Nay	Aye	Aye	假
Aye	Nay	Nay	假
Aye	Nay	Aye	假
Aye	Aye	Nay	假
Aye	Aye	Aye	真

图 6-10 显示了可以完美地分类训练数据的树，它需要五个内部拆分结点。不用关心这棵树是如何创建的，因为对我们而言那不重要。

允许使用逻辑和函数组合不同特征，形成逻辑连接可以产生图 6-11 中简单得多的树。第二棵树阐明了使用字段组合带来的另一个潜在好处，该树现在较能准确表达类全体一致性的概念："当所有投票者赞同时，决策是全体一致的。"

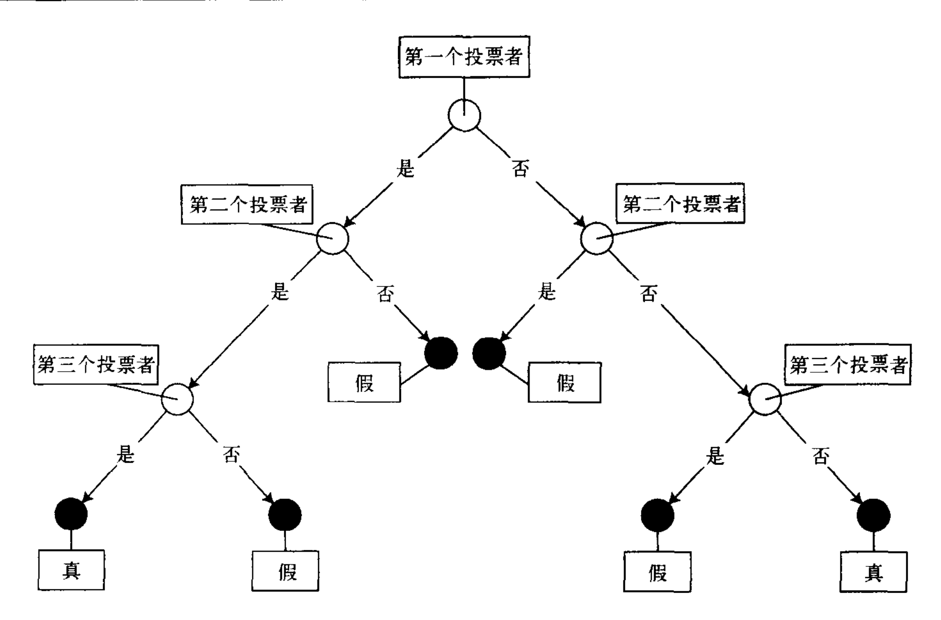


图 6-10 一致性函数按单个字段拆分的最佳二元树

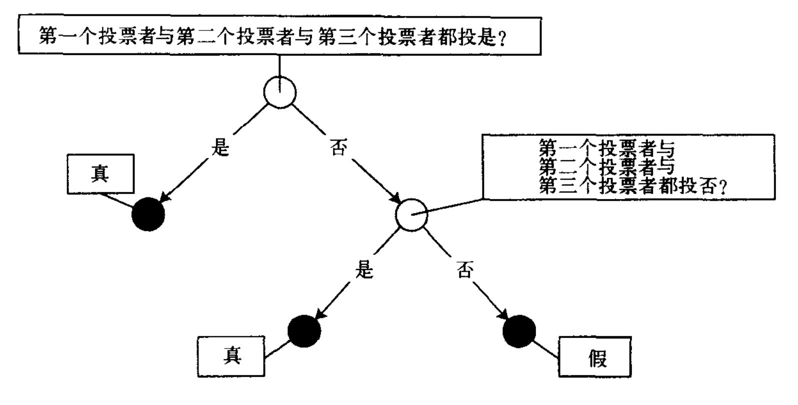


图 6-11 组合特征简化了定义一致性的树

一棵一看就懂的树，被机器学习研究者表述为具有好的“思维适应”。在机器学习领域的一些研究者重视这一观念，.但它似乎是一个他们围绕其建立研究的微小的、结构良好问题的人工制品。在现实世界中，如果一个分类任务如此简单，以至于你能围绕代表它的整个决策树得到你的主意，大概就不需要浪费时间使用强有力的数据挖掘工具去发现它了。我们相信理解通向任何特定叶的规则的能力是很重要的，相反，在实验室之外，瞟一眼就能说明整个决策树的能力既不重要，也似乎不太可能。

6.7.2 倾斜超平面

分类问题有时可以用几何术语表示，这种思考方法对于所有字段具有连续变量的数据集尤其有用。在这一思考方式中，每条记录是在多维空间中的一个点，每个字段代表记录在该空间中沿着某个轴的位置。决策树是把空间切分为区域的一种方式，每个区域被标记为一个类，任何落入其中一个区域的新记录被归入相应的类。

在每个结点上测试单个字段值的传统决策树只能形成矩形（rectangular）区域。在一个二维空间内，公式“Y 小于某一常量”的测试形成以垂直于 Y 轴且平行于 X 轴的直线为边界的区域。不同的常量值会使该直线向上和向下移动，但该直线保持水平。类似地，在一个较高维的空间内，对单个字段进行的测试定义了一个“垂直于代表测试中使用字段的轴，并且平行于所有其他轴的超平面”。在二维空间内，只有水平线和垂直线起作用，产生的区域是矩形的；在三维空间内，相应的形状是长方体；在多维空间内，就是超矩形。

问题是有些事物并不恰好适应矩形方格。图 6-12 说明了这个问题：两个区域实际是按对角线划分的，它需要一棵很深的树以产生足够的矩形来近似地表示。

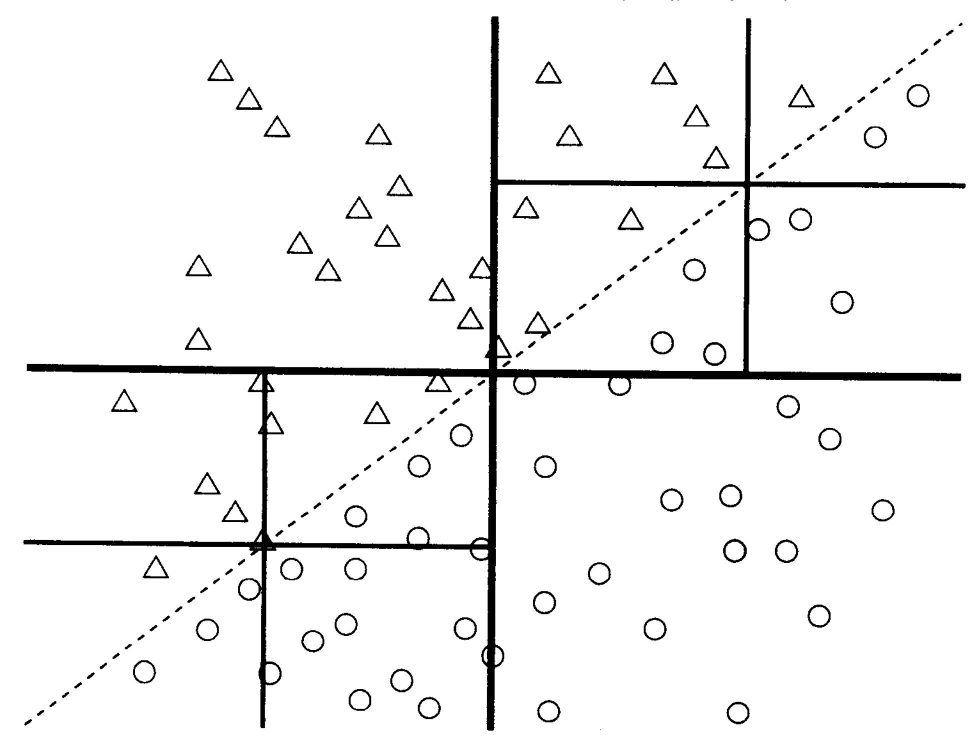


图 6-12　左上和右下象限很容易分类，而其他两个象限必须在区域之间划分为许多小的方格作为近似边界

在这种情况下，正确的解决办法很容易发现，那就是允许把要考虑的属性进行线形组合，一些软件包企图通过基于字段值的加权和拆分来倾斜超平面。通常有多种爬山方法可用于选择权重。

当然，即使允许取对角线，也很容易遇到那些不易被捕捉到的区域。这种区域可能有弯曲边界，并且字段可能不得不用更复杂的方式组合起来（例如长乘宽得到面积）等。在建树过程中，除了给建树过程仔细选择输入字段，在必要时创建用于捕捉被该领域专家已知或可疑的关系的衍生字段以外，没有什么可以代替的方法。这些衍生字段可以是若干其他字段的函数，像自动组合字段倾斜超平面一样，手工插入的衍生字段针对同样的目的。

6.7.3 神经树

在每个结点组合许多字段输入的一种方式是使每个结点包含一个小的神经网络。对于用

矩形区域不能很好描述真实分类形状的领域，神经树能产生更加精确的分类，而且比纯的神经网络能更快地进行训练并给出得分。

从用户的观点看，与决策树的变体相比，这一杂化技术与神经网络变体有更多的共同点。因为它与其他神经网络技术相同，没有能力解释做出的决策。该树仍旧会产生规则，但都是按 $F(w1x1, w2x2, w3x3, \cdots) \leqslant N$ 的形式，其中 F 是神经网络使用的组合函数。这种规则对神经网络软件比对人更有意义。

6.7.4　使用树分段回归

另一个把树和其他建模方法组合起来的例子是分段线性回归的形式，在其中，决策树中的每个拆分被选中的目的，就是使得在该结点上数据的简单回归模型的误差最小化，同样的方法可以应用于分类型目标变量的逻辑回归。

6.8　决策树的替代表示法

传统的树图形是表示决策树实际结构的很有效方式。当焦点更关注于结点的相对大小和集中度时，其他表示法有时更有用。

6.8.1　方格图

尽管“树图”和“二十问题”类推对于决策树方法中某些性质的形象化是有帮助的，但在某些情形下，方格图（box diagram）可能更加直观。图 6-13 显示了决策树的方格图表示法，它试图基于年龄和最近看过的电影把人们分类为男性或女性，该图可视为一种二维散点图的嵌套集合。

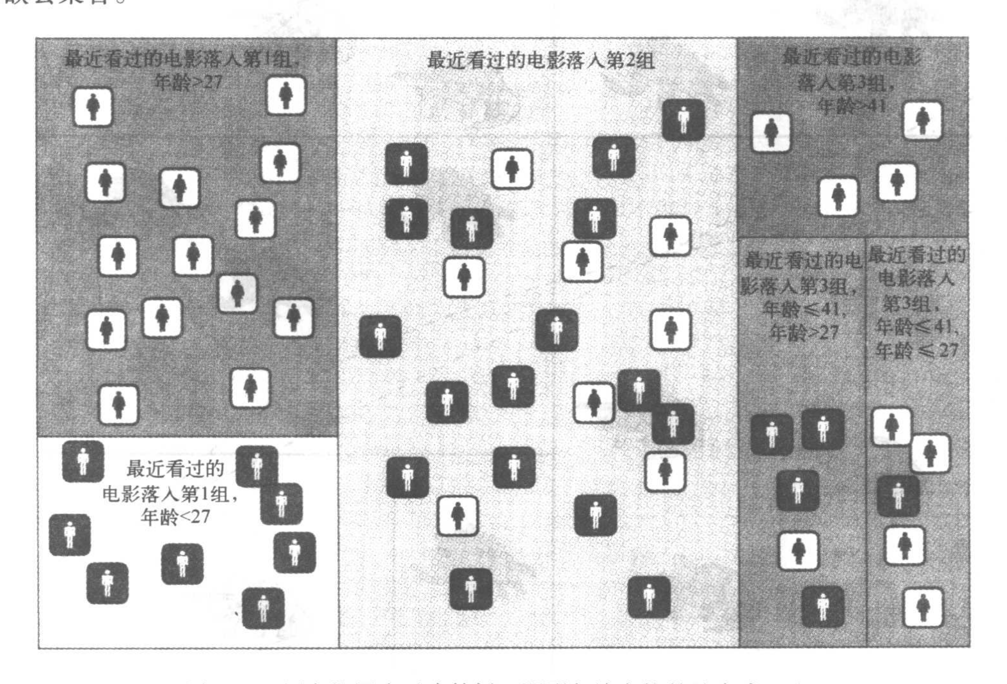


图 6-13　用方格图表示决策树。阴影与该方格的纯度成正比；大小与其中的记录数目成正比

在决策树的根结点上，最初三路拆分是基于被调查者回答的最近看过的电影归入三组中的哪一个进行拆分的。在该图最外面的方格中，水平轴就代表那个字段。最外面的方格被分成小块，每一小块代表树中下一层的某个结点。每一小块的大小与落入其中的记录数成正比。其次，每个方格的垂直轴用于表示作为那个结点下一拆分器的字段。一般对每个方格而言，这将是一个不同的字段。

现在有一个新的方格集合，其中每个方格代表树的第三层的一个结点。继续这一过程，一直把方格分割到树的每个叶，使它们有自己的方格。由于决策树深度常常并不一致，所以一些方格可以比其他方格更频繁地被再分。在一个二维图表中，方格图更容易表示分类规则，因为这些规则依赖于图中任意一个变量的数值。

由此得出的图很有表现力。当我们向表格中投掷记录时，它们落入特定的方格并归入相应分类。方格图允许我们在若干层次的详细程度上观察数据。扫一眼图 6-13 就可以看出左侧底部包含了高集中度的男性。

更仔细地看，我们发现一些方格在分类或收集大量记录方面似乎做得特别好。按照这种观察，把决策树看做围绕相似点的群组绘制方格的一种方法就是很自然的。所有在特定方格中的点按相同的方式分类，因为它们都满足定义那个方格的规则。这与通过画直线或椭圆曲线穿过数据间隔试图把数据分割为类的传统统计学分类方法（如线性、对数和二次方程式判别）形成鲜明的对比。两类方法的基本区别是：当一条记录有若干不同的方式划为目标类的一部分时，使用单条线来发现类别之间边界的统计学方法是软弱无力的。图 6-14 使用两种恐龙阐明了这一点，决策树（表示为方格图）成功地从三角恐龙中独立出剑龙。

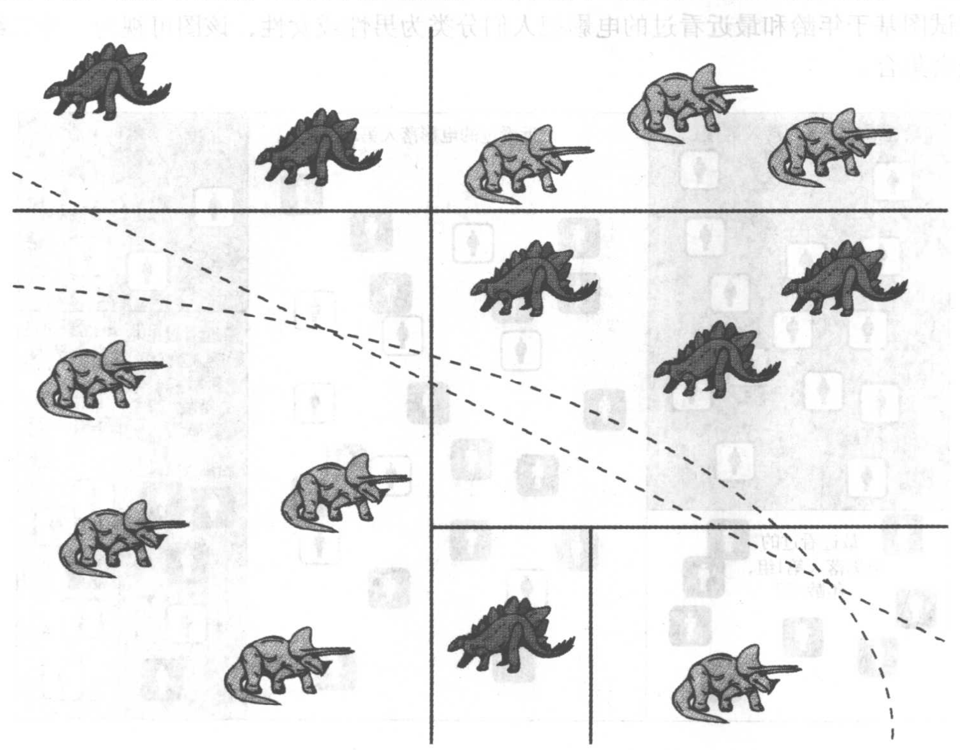


图 6-14　一条简单的直线或曲线常常不能分割不同的区域，而决策树做得更好

例如，在信用卡行业中，客户可以有若干方式给公司带来利润：一些有利可图的客户具有很低的交易率，但一直保持没有拖欠的高周转余额；另一些虽然每个月全部花光他们账户中的所有余额，但仍然有利可图，因为他们产生了高交易量；还有其他一些有很少的几笔交易，但偶尔进行大额采购并花几个月付清欠款。两个非常相异的客户也许带来同等利润。决策树能够发现每个单独的群组，对它做出标记，并通过提供方格本身的描述提示每个群组收益的原因。

6.8.2 树年轮图

来自 SAS 研究所的 Enterprise Miner 产品使用了决策树的另一种巧妙表示方法。图 6-15 中的图表看起来像是树已被砍倒，我们来查看树桩。

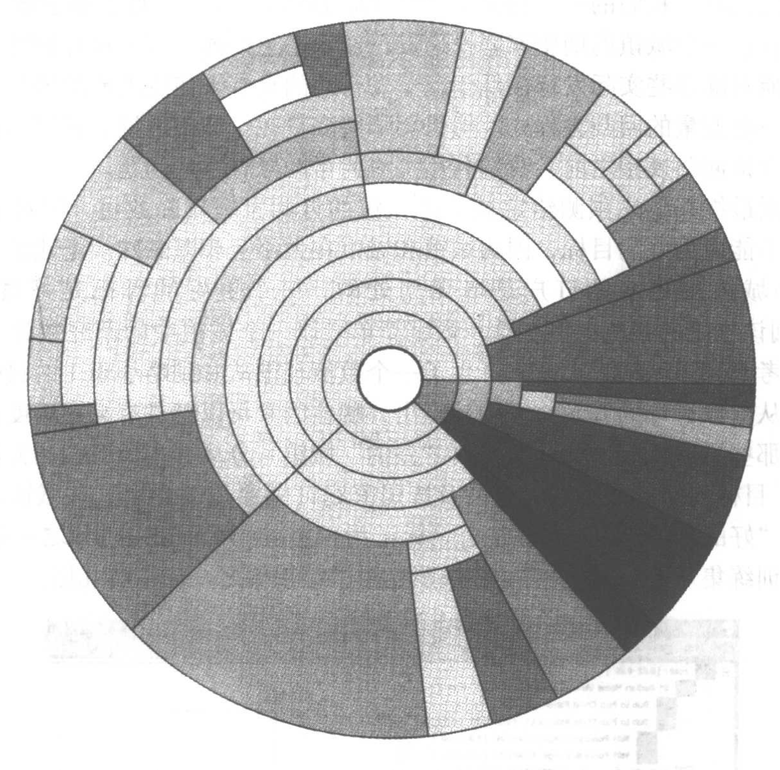


图 6-15 用 SAS Enterprise Miner 做出的树年轮图汇总了树的不同层

图中央的圆圈表示做出任何拆分之前的根结点。从中央向外移动，每个同心圆环表示树的一个新的层，最接近中央的环表示根结点拆分。两条路径之间的弧长与其中的记录数目成正比，阴影代表该结点的纯度。在本图所示模型中的首次拆分是相当不均衡的。它把记录划分为两个群组——集中度与父总体没有多大差别的一个大群组，加上一个具有很高的目标类集中度的小群组。在下一层上，这一方块点再次拆分，那个用细的、一路延伸到图表最外层环的暗扇形区表示的一个分支就是一个叶结点。

该年轮图一目了然地显示了树的深度和复杂性，并指示出在目标类上有高集中度的位置。它没有直接显示定义这些结点的规则，用户点击图表的特定区域时，该软件会把这些规则展现出来。

6.9 实际应用中的决策树

决策树能够应用于许多不同的情形：

- 探查大的数据集以挑选出有用的变量；
- 预测工业过程中重要变量的未来状态；
- 为推荐系统形成指导性的客户簇。

本节包含了用于以上这些情形的决策树的示例。

6.9.1 决策树作为数据探查工具

在数据挖掘项目的数据探查阶段，要挑选出那些对预测特定目标可能重要的变量，决策树是一个有用的工具。我们的一个报业客户，*The Boston Globe*，对于基于多种人口统计学和地理特征，评估一个城镇的期望家庭投递发行额水平感兴趣。有了这样的评估，在其他事务中，他们将能对准那些实际发行额低于期望发行额的具有未使用潜能的城镇。最终的模型将是一个基于一些变量的回归方程式。但采用哪些变量呢？准确地说，该回归将试图评估哪些内容？在建立该回归模型之前，我们利用决策树帮助探查这些问题。

尽管该报纸最终兴趣是预测给定城市或城镇的订阅家庭实际数目，但对于一个回归模型，该数目并不能成为好的目标，因为城镇和城市在大小上相差悬殊。把建模力量浪费在发现大城镇比小城镇有更多的订户是毫无用处的。一个更好的目标是穿透度（penetration）——订阅该报纸的家庭比例，这一数字简单乘以一个城镇中的家庭数就可生成订户总数估计。还要考虑到，城镇大小因素产生了一个数值范围从 0 到略小于 1 的目标变量。

下一步是从数以百计的城镇特征中计算出，哪些因素可以把具有高穿透度的城镇（“好的”城镇）从那些具有低穿透度的城镇（“差的”城镇）分离开来。我们的方法是建立具有二元“好/差”目标变量的决策树。这涉及按照家庭订阅穿透度排序这些城镇，并把顶部三分之一标记为“好的”，把底部的三分之一标记为“差的”，而中间三分之一是好是差不明朗——不计入训练集。图 6-16 中的屏幕快照显示了结果树之一的顶端几层。

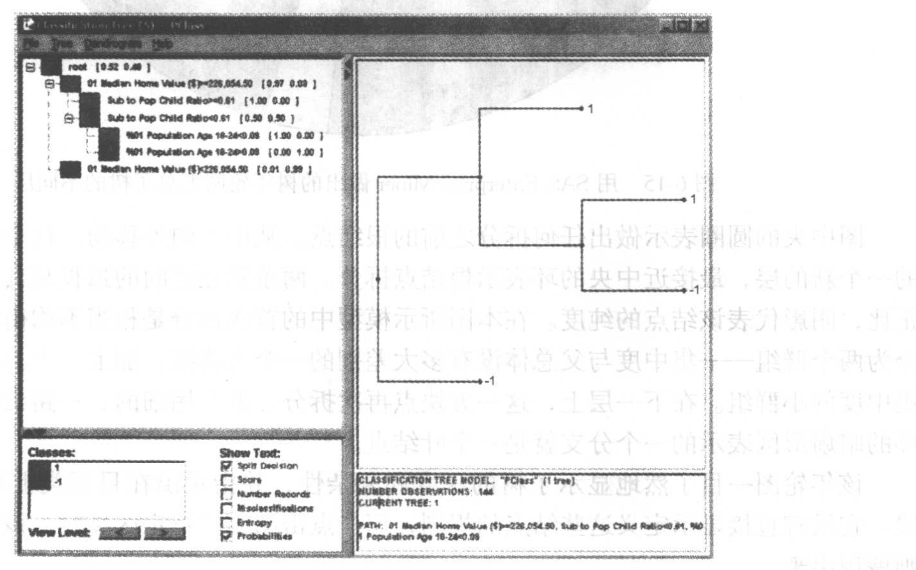


图 6-16 一个把好的城镇与差的城镇分离开来的决策树，正如 Insightful Miner 工具所示

该树显示中值住宅价值是最好的首次拆分，中值住宅价值（在该郡中具有一些最昂贵的房屋的区域）小于226 000美元的城镇对于这份报纸是不良的潜在客户。下一层的拆分更令人惊讶，为拆分选中的变量是把城镇中的订户与该城镇人口整体相比较得到的一族衍生变量(derived variable）中的一个。按照家庭投递穿透度、订户与总人口情况相近的城镇比那些订户远离均值的城镇更好。可以把好的与差的城镇区分开来的其他重要变量，包括学校就读的平均年数、在蓝领职业中人口的百分比和在高级职业中人口的百分比，所有这些最终都被作为该回归模型的输入值。

我们曾经期望会有一些其他重要的变量，例如离波士顿的距离和家庭收入，结果被证明是不太有力的。一旦决策树把焦点集中于一个要么包含进来，要么不被使用的变量，其中原因常常稍稍一想就清楚。例如，与波士顿距离的问题是，当一个人首次驾车外出到城郊时，家庭穿透度随着离波士顿的距离上升；然而过不了多久，当人们远离波士顿，不再那么多地关心那个地方发生的事情时，离波士顿的距离与穿透度变为负相关。住宅价格是一个好的预测器，因为它的分布类似于目标变量的分布，在最初几英里增加随后下降。决策树不仅指导我们思考什么，也可指导我们使用哪些变量。

6.9.2 把决策树方法应用于顺序事件

预测未来是数据挖掘最重要的应用之一。在历史数据中分析趋势，以便预测未来行为，这种任务在我们已经测试的每一个领域经常重复出现。

我们的客户之一，一家大银行，为了找准在其支票账户中的预警信号，仔细考察了客户的详细交易数据。随时跟踪自动柜员机提款、直接工资单存款、余额查询、直接柜台存取以及数以百计客户流失的其他交易类型和客户属性，以发现那些能让银行识别出顾客忠诚度开始变弱但仍有足够时间采取正确行动进行挽回的特征。

另一个客户，一家内燃机制造商，使用SPSS的Clementine数据挖掘套件中的决策树组件，以卡车登记的历史数据为基础，预测内燃机销售，目的是识别出那些可能准备对其大装备的发动机进行更换的人们。

销售、利润、失效方式、流行趋势、物品价格、运行温度、利率、呼叫音量、响应率和返回率，所有这些都是人们试图预测的内容。在某些领域，尤其是经济学领域，时间序列数据的分析是统计分析的当务之急，因而你可能指望存在一系列现成的可用技术，能够用于时间顺序数据上的预言性数据挖掘，不幸的是，实际情况并不是这样的。

首先，许多其他领域的时间序列分析工作集中于分析单个变量中的模式，例如孤立的美元兑日元的汇率或失业率；公司的数据仓库有可能包含展示周期模式的数据。当然，支票账户中的平均日余额反映出租金通常是在每月初支付的，并且许多人是在星期五支付租金。但是，在极大程度上，这些种类的模式并不令人感兴趣，因为它们既不是意外的又不是可采取行动的。

在商业数据挖掘中，我们更关注的是，如何把大量数目的单独变量组合起来，以预测某些未来的结果。第9章会讨论时间如何被整合到关联规则（association rule）中以便发现序列模式（sequential pattern)。决策树方法在这一领域也被成功地应用，但它一般需要结合趋势信息，通过包括诸如明确表示随时间改变的变化差值和比率等字段来丰富数据。第17章中会更详尽地讨论这些数据准备问题。下一小节描述了一个应用，即自动产生这些衍生字

段，并利用它们建立一个基于树的模拟器，通过该模拟器投影出一个未来的完整数据库。

6.9.3 模拟未来

这一讨论很大程度上基于 Marc Goodman 的讨论和他的 1995 年关于一项称为投射可视化技术的博士论文。投射可视化使用一个历史数据快照的数据库来开发模拟器，运行该模拟器能够投射出所有变量将来的数值。最后的结果是得到一个扩展的数据库，其中的新记录具有和原始记录完全相同的字段，但其值是由该模拟器给出的，而不是观察得到的，这种方法在后面的“使用决策树用于投射可视化”部分会更详细地描述。

案例研究：咖啡烘烤工厂的过程控制

雀巢，世界上最大的食品和饮料公司之一，使用大量的连续进料咖啡烘烤机生产多种咖啡产品，包括 Nescafé Granules、Gold Blend、Gold Blend Decaf 和 Blend37 等。其每一种产品有一个“配方”，规定了一系列烘烤机变量的目标数值，例如在各种排气点的空气温度、各种风扇的速度、气体燃烧的比率、使咖啡豆淬火要导入的水量和各种风门片和阀门的位置。当烘烤咖啡时，有许多情况可能使事情变糟，从烘烤成色太浅到代价很大并损害烘烤机的起火，这些情况都有可能发生。一批烘烤不好的咖啡会导致付出很大的代价，而对设备的损害代价则更加昂贵。

为帮助操作者保持烘烤机恰当地运转，需要从大约 60 个传感器中收集数据，每 30 秒钟，这一数据连同控制信息被写入日志，并以图形形式表示，使其对操作者直观可用。这里描述的工程发生在英国约克郡的一家雀巢研究实验室，雀巢利用投射可视化，以传感器日志为基础，建立了一个咖啡烘烤机模拟器。

模拟器的目标

雀巢发现咖啡烘烤机模拟器能以若干方式改善其生产过程。

- 通过使用模拟器来实验新的配方，不用中断生产就能够评价大量新配方。而且，能够提前排除可能会导致烘烤机着火或其他损坏性的配方。
- 模拟器能够用于训练新操作者，并模拟他们可能遇到的常规问题，教给他们相应的解决方案。使用模拟器，操作者可以尝试用不同的方法来解决问题。
- 模拟器能够跟踪实际烘烤机的操作，并向未来投影几分钟。当模拟器遇到一个问题，能够给出警报，操作者有足够的时间解决麻烦。

使用决策树进行投射可视化

套用机器学习领域的 Goodman 术语，每一个瞬间的快照被称为一个情形。情形由属性组成，属性就是情形记录中的字段。属性可以是任何数据类型，可以是连续型也可以是分类型。属性用于形成特征，特征是用多种方式组合起来形成决策树内在结点的布尔型（是/否）变量。例如，如果数据库包含一个数值型的薪金字段，这是一个连续属性，那么这可能导致创建一个诸如薪金<38 500 的特征。

对于像薪金这样的连续变量，要为训练集中观察到的每个值产生“属性≤值”这种形式的特征，这意味着可能存在与训练集中情形一样多的衍生于属性的特征。基于等同性或集合成员的特征就会按照符号属性和诸如人名或场所的文字属性生成。

属性也可以用于生成解释，这些属性指的是衍生于给定属性的新属性，而解释通常反映该领域的知识以及哪种关系可能是重要的。在当前的问题中，发现属性的值从一个时段到另

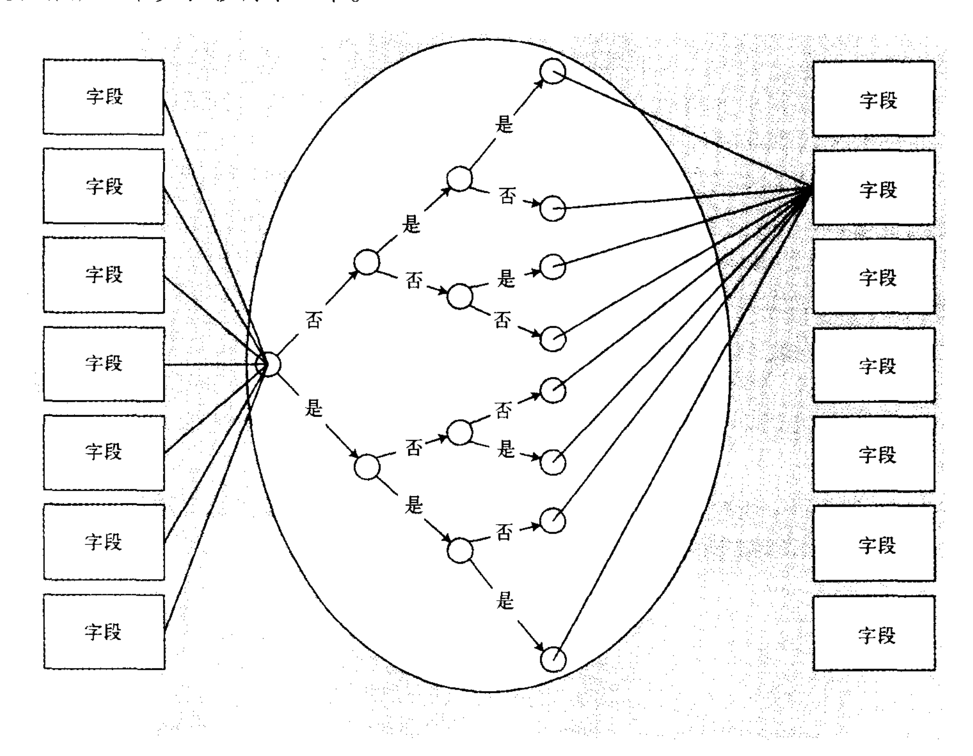
一个时段，其数量、方向和变化率随时间变化的模式可能是重要的。因此，对每个数值属性，软件自动生成对于属性差异以及属性首个离散值和第二个导出值的解释。

然而，一般而言，用户会提供解释。例如，在一个信用风险模型中，债务与收入的比率可能会比它们的大小更具有预言性。利用这一知识，我们可能添加一个关于这两个属性比率的解释。通常用户提供的解释以程序不能自动给出的方式与属性组合，具体的例子包括从纬度和经度变化计算大圆周距离，或者利用三种线性度量的乘积得到体积。

从一个情形到下一个情形

投射可视化背后的中心思想是利用历史的情形产生一套规则，可以用于从情形 n 生成情形 $n+1$。当这一模型应用于最终的观察情形时，它产生一个新的投射情形。要投射出将来的多于一个时间步长的情形，我们可以继续将该模型应用于最近创建的情形。当然，随着该模拟器运行越来越多的时间步长，投射出的数值置信度会下降。

下图显示了一个单一属性的投影方式，使用以前一个情形为基础从所有其他属性和解释生成特征的决策树。在训练过程中，对每一属性生成一个独立的决策树。整个森林被评价，以便模拟器从一个步长移向下一个。



快照使用决策树创建下一时间快照

烘烤机模拟器的评价

该模拟器是使用包含 34 000 个情形的训练集建立的，然后该模拟器用包含不属于训练集组成部分的大约 40 000 个其他情形的测试集进行评价。对于测试集中的每个情形，模拟器产生深入未来 60 步长的投射快照。在每一步长上，所有变量的投射数值都与实际值进行比较。正如预期的那样，误差大小随时间增加。例如，投射的每分钟产品温度的误差率是

2/3℃，但即使投射深入未来30分钟，该模拟器的工作也比随意猜测好很多。

投射趋势的结果证明，除了最有经验的少数操作者外，烘烤机模拟器比所有其他操作者都更准确，当然，即使最有经验的操作者，有了模拟器的帮助也能做得更好。操作者乐于使用该模拟器，并报告说它给了他们对正确行动的新的洞察力。

6.10 小结

决策树方法对数据探查、分类和评分等方法具有广泛的适应性。它们也能用于估计连续值，尽管很少是首选方案，因为决策树产生“成块的”估计——到达同一个叶的所有记录被赋予相同的评估值。当数据挖掘任务是分类记录或预测离散结果时，决策树是首选对象。当你的目标是把每个记录分类到少数宽泛范畴之一时，可以选择使用决策树。理论上，决策树能够把记录分配到任意数目的类中，但在每个类的训练样本数目变小时它们容易出错，这种情况可能在具有许多层和（或）每个结点具有许多分支的树中更容易发生。在许多商业情形中，问题会自然分解到诸如“响应者/非响应者”或者“好/坏”的二元分类，因此在实践中这不是大问题。

当目标是生成可理解的和可解释的规则时，决策树也是一个很自然的选择。决策树技术最强大的功能之一是，可以产生能被翻译成可理解的自然语言或SQL的能力，即便在复杂的决策树中，通常也非常容易通过树追踪任何一条路径以到达某个特定的叶，因此它对任何特定分类或预测的解释相对比较简明。

决策树比许多其他技术需要的数据准备更少，因为它们既擅长处理连续型变量，也擅长处理分类型变量。作为神经网络和统计学技术难题的分类变量，通过形成不同类别群组加以拆分，而连续变量通过划分值的范围进行拆分。因为决策树不使用数值变量的实际数值，因此它们对离群值和不均匀的分布不敏感。这些强有力的适应性是以丢弃一些在训练集中有用的信息为代价的，因此调试良好的神经网络或回归模型常常比决策树更能利用好相同的字段。正因为这一原因，决策树经常用于挑选出一组好的变量用作另一建模技术的输入变量。面向时间的数据确实需要大量数据准备工作。时间序列数据必须被强化，这样趋势和序列模式才会变得清晰可见。

决策树会给出关于所应用的数据的许多信息，所以作者大概在每一数据挖掘工程中都会用到决策树，即便最终模型将使用某些其他技术创建。

第 7 章　人工神经网络

在许多数据挖掘和决策支持应用中，由于有公认的轨迹记录，人工神经网络已经成为一种普遍采用的方法。神经网络（“人工”二字通常被省略）是一种可以容易地应用于预测、分类和聚类的强有力的多用途工具。从预测金融业的时间序列到医学诊断情形，从识别有价值的客户群到识别欺诈信用卡交易，从识别写在支票上的数字到预测发动机的故障率，等等，人工神经网络被广泛应用到各行各业。

当然，最有力的神经网络是生物所具有的那些神经网络，人的大脑使人们能够总结经验。与此对应的是，计算机则通常擅长于反复执行明确的指令。神经网络的魅力在于，通过在数字计算机上模拟人脑的神经联系，桥连二者之间的隔阂。在明确定义的领域中使用时，从某种意义上来说，神经网络从数据中概括和学习的能力，是模仿我们从经验中学习的能力。这种能力对数据挖掘是有用的，而且也使神经网络的研究成为令人兴奋的领域，预示着未来有新的和更好的结果。

但一个缺点是，训练神经网络所得到的结果是遍及网络内部的权重，这些权重所提供的关于“为什么解决方案正确”的洞察，一点也不比剖析人脑以解释人的思维过程的方法所能提供的更多。也许将来有一天，探查神经网络的尖端技术可以帮助提供一些解释。同时，像人类大脑的工作一样神秘，神经网络也是通过内部运行的暗箱很好地完成它的工作，如同古希腊人顶礼膜拜的 Delphi 神所示的回应一样，神经网络产生的答案时常是正确的。这些答案具有商业价值——在许多情况下，是一个比提供解释更重要的特征。

本章将首先回顾一下历史，神经网络起源于通过在计算机上建立模型来模仿人类智能的实际尝试。然后，在开始分析技术性细节之前，讨论回顾了使用这项技术进行房地产评估的早期案例历史。本章的大部分内容把神经网络作为预言性建模工具，在本章的最后，讲述如何将它们应用于非定向数据挖掘。一如既往，我们还是从讲述历史开始。

7.1　历史回眸

在计算机科学年鉴上，神经网络有耐人寻味的历史。关于神经元功能的最初研究——生物神经元——出现在 20 世纪 30 年代和 40 年代，比数字计算机产生的历史还早。1943 年，耶鲁大学的神经生理学家 Warren McCulloch 和逻辑学家 Walter Pitts 设计了一个简单的模型，解释生物神经元如何工作，撰写并出版了题为“神经活动中内在的逻辑运算法”的论文。尽管他们关注的目标是理解大脑解剖结构，最后的结果却是：该模型给人工智能领域提供了某种灵感，并且最终提供了一种解决神经生物学以外的特定问题的新方法。

在 20 世纪 50 年代，当数字计算机开始出现时，以 McCulloch 和 Pitts 的工作为基础，计算机科学家设计出了被称为感知器（perceptron）的模型。这些早期的网络解决的问题实例是，如何通过来回控制手推车运动来平衡竖立在手推车上的扫帚。当扫帚开始向左侧歪倒时，手推车随之向左移动使它保持直立。虽然在实验室里，出现了有限的少数几个使用感知器并获得成功的例子，但作为解决问题的普通方法，其结果却令人失望。

早期神经网络应用受限的一个原因是，在那个时代，功能最强的计算机也比不上今天廉

价的台式计算机。另一个理由是，就像1968年Seymour Papert和Marvin Minsky（马萨诸塞工学院的两位教授）所揭示的，这些简单的网络有理论缺陷。由于这些原因，在20世纪70年代，神经网络在计算机上实现的研究大幅减缓。后来，加州工学院的John Hopfield在1982年发明了反向传播（back propagation），一种避开较早方法的理论缺陷的神经网络训练方法。这一发展引发了神经网络研究的复兴。在整个20世纪80年代，研究从实验室转向商业界。此后研究被用于既解决操作性问题（例如探测欺诈信用卡交易的发生，识别支票上所写金额），又解决数据挖掘面临的挑战。

在人工智能研究人员开发类似于生物活动模型的神经网络的同时，统计学家正在利用计算机，扩展统计方法的能力，一种被称为逻辑回归技术，被证明对许多统计分析特别有价值。如同线性回归一样，逻辑回归试着画出一条适应观察数据的曲线。它不使用直线，而是使用一种函数，被称为逻辑函数。逻辑回归以及其同类线性回归都可以看做神经网络的特例。事实上，神经网络的全部理论都可以使用统计方法来解释，如概率分布、可能性等。然而，出于解释的目的，本章更多地倾向于生物模型，而不是纯理论的统计学。

受多种因素的影响，神经网络在20世纪80年代变得很受欢迎。首先，计算能力完全能满足要求，尤其是在商业领域中有众多的数据可以利用；其次，由于分析人员认识到它们与已知的统计方法密切相关，使用神经网络变得更加得心应手；第三，由于多数公司的操作系统已经实现自动化，所以有相关的数据；第四，实际的应用比纯粹的人工智能方法更重要，帮助人们构建的工具已经取代了制造假人的目标。由于具备确凿无疑的功能，神经网络现在已经是（并将继续是）数据挖掘非常受欢迎的工具。

7.2 房地产评估

与人类专家从经验获取知识的方法完全相同，神经网络有能力通过案例学习。下面的例子应用神经网络解决多数读者熟悉的问题——房地产评估。

为什么要进行自动评估呢？很清楚，自动评估可以帮助房地产代理商较好地将预期的买主与预期住宅匹配到一起，改进经验不足的代理商的生产率。另外一种用途是建立资讯服务站或Web页，预期的买主可以在那里描述想要的住宅——而且直接得到关于他们梦寐以求的房子需要多少花费的反馈。

也许意想不到的应用是在二级抵押市场上。因为影响拖欠的主要因素是风险财产价值的比例。好的、协调一致的评估对评定个人贷款和贷款组合的风险相当重要。如果贷款量超过100%的市场价值，拖欠的风险上升得相当快，一旦给出贷款，市场价值如何计算？为弄清这个意图，美国联邦住宅抵押贷款公司（Federal Home Loan Mortgage Corporation）的Freddie Mac开发了被称为贷款勘探者（Loan Prospector）的一种产品，对美国各处住宅自动地进行评估。贷款勘探者最初是由圣地亚哥的一家公司HNC基于神经网络技术开发的，该公司现已并入Fair Isaac公司。

回到刚才的实例，该神经网络模仿评估师，根据财产特征估计住宅市场价格（见图7-1）。她知道在城镇的某一个区域中的住宅比在其他地区昂贵。附加的卧室、较大的车库、住宅的风格和占地面积的大小是要考虑的其他因素。她没有使用一些固定的方法，而是在平衡类似销售房价的经验和知识，而且她所拥有的购房价格知识不是静态的。她知道整个区域近期住宅的销售价格，并且能识别随时间变化的价格趋势——为适应最近的数据精心调整测算。

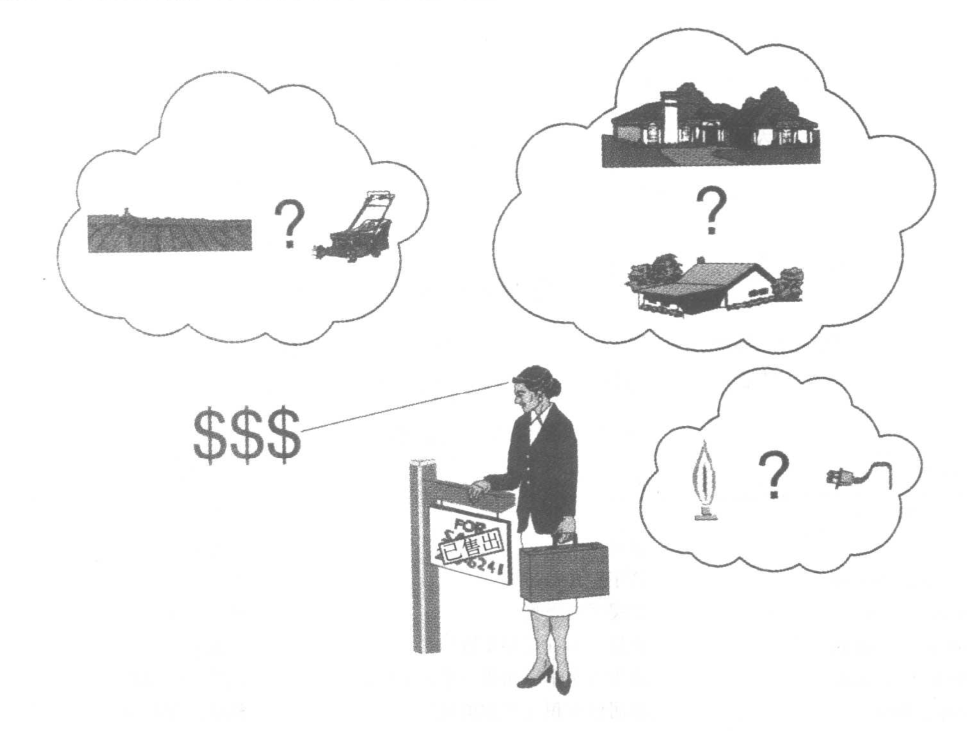


图 7-1　房地产代理商和评估师结合住宅的特征，提出评估值——生物神经网络工作实例

在一个明确定义的领域中，评估师或房地产代理商是人类专家的很好示例。经专家考虑的一套固定标准用以描述住宅的特征，并且变成评估值。IBM 的研究人员在 1992 年承认这是一个好的神经网络问题，图 7-2 说明了为什么。神经网络获得特定的输入——在本例中是来自房产相关单据的信息——并将它们变为特定的输出，即住宅的评估值。因为以下两个因素，输入的列表是明确定义的：一是扩充使用多重列表服务（MLS）来共享不同的房地产代理商的房屋资讯市场信息；二是将二级市场上出售的抵押住宅描述标准化。期望的输出也是明确定义的——特定的美元数目。除此之外，在早期的销售形式中有很丰富的经验，可以教会网络如何评估住宅。

提示： 神经网络非常适合于预测和估计问题。好的问题有下列三个特点：

- 输入很容易理解。你非常清楚数据的哪些特征是很重要的，但是没有必要知道如何将它们结合在一起。
- 输出很容易理解。你知道正在尝试建立的模型的内容。
- 经验是非常有用的。你有大量的输入和输出是已知的例子，用这些案例来训练网络。

建立神经网络以便计算出估计的住宅价值，第一步就是确定一组影响销售价格的特征，可能的常见特征见表 7-1。实际上，这些特征只对某个地理区域的住宅起作用。为了扩展评价实例以处理许多邻近地区住宅的估价，输入数据可能还要包括邮政编码信息、邻近地区的人口统计信息和其他邻近地区生活质量指标，例如学校排名和交通方便程度。为了简化实例，这些附加特征不包括在我们的讨论中。

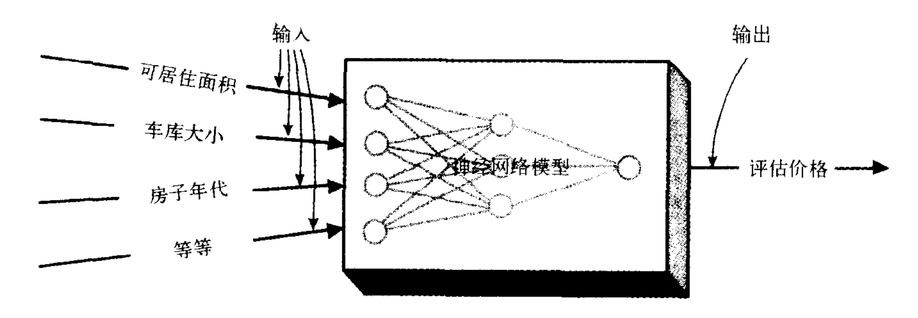


图 7-2 神经网络就像一个知道如何处理输入以产生输出的黑匣子，计算相当复杂且难以理解，但给出的结果时常是有用的

表 7-1 描述住宅的普通特征

特征字段	描 述	取值范围
Num_Apartments	住宅单元数目	整数：1～3
Year_Built	建造年代	整数：1850～1986
Plumbing_Fixtures	管子附件数目	整数：5～17
Heating_Type	取暖系统类型	代码：A 或者 B
Basement_Garage	底层车库（可存车数）	整数：0～2
Attached_Garage	附带木制车库面积（平方英尺）	整数：0～228
Living_Area	总居住面积（平方英尺）	整数：714～4185
Deck_Area	露台/敞开门廊面积（平方英尺）	整数：0～738
Porch_Area	室内门廊面积（平方英尺）	整数：0～452
Recroom_Area	娱乐室面积（平方英尺）	整数：0～672
Basement_Area	已完成地下室面积（平方英尺）	整数：0～810

训练网络，建立估计模型，然后将模型用于估计未知实例的目标价值。将已知的实例（来自以前销售的数据）用于训练网络，使它学会如何计算销售价格。训练实例需要有另外两个附加的特征：房子的销售价格和销售日期。销售价格被用作目标变量，而日期则用于将实例分割成训练集、验证集和测试集，表 7-2 显示的是来自训练集的一个实例。

表 7-2 来自训练集的样本记录，价值按比例换算到（−1，1）的范围

特征字段	取值范围	原始值	比例映射值
Sales_Price	\$103 000 − \$250 000	\$171 000	−0.0748
Months_Ago	0～23	4	−0.6522
Num_Apartments	1～3	1	−1.0000
Year_Built	1850～1986	1923	+0.0730
Plumbing_Fixtures	5～17	9	−0.3077
Heating_Type	代码为 A 或 B	B	+1.0000
Basement_Garage	0～2	0	−1.0000
Attached_Garage	0～228	120	+0.0524
Living_Area	714～4185	1614	−0.4813
Deck_Area	0～738	0	−1.0000
Porch_Area	0～452	210	−0.0706
Recroom_Area	0～672	0	−1.0000
Basement_Area	0～810	175	−0.5672

训练网络的过程实际上是内部调整权重的过程，以便最终达到权重的最佳组合，实现期望的预测。神经网络从随机的一组权重开始，它最初的表现很差，但通过在训练集上反复进行训练，并且每次调整内部权重以减小总误差，网络在训练集上的表现会越来越好，逐渐地接近目标数值，直到近似值不再改变时，网络停止训练。

这个调整权重的过程对进入数据的表示法比较敏感。例如，考虑数据中用于测定占地面积大小的字段，如果占地面积大小按英亩计量，则可能的合理值大约在 1/8 英亩到 1 英亩之间；如果按平方英尺测量，同一面积的数值可能是从 5 445 平方英尺到 43 560 平方英尺。然而，由于技术原因，神经网络把输入限制为小的数值，比如 -1 和 1 之间。举例来说，如果一个输入变量具有比其他输入变量大很多的值，这个变量就会在目标变量的计算过程中占有优势，神经网络就会消耗宝贵的迭代来减小这一输入的权重，以减小它对输出的影响。即，网络将会找到的第一种“模式”是，占地面积变量的数值较其他变量的值大得多。既然这不是特别有意义的，最好还是使用英亩而不是平方英尺测定占地面积大小。

这个思想概括为，神经网络的输入通常应该是小一点的数值，最好将其限制在小范围内，如 -1 到 1，而这需要在训练网络之前映射所有的数值（像绘制地图那样按比例变换数值），包括连续型和分类型数值。

映射连续数值的一种方法是把它们变成分数，一般是用该值减去数值范围的中值，将结果除以范围的大小，然后乘以 2。例如，为得到建造年代 Year_Built（1923）的映射值，从 1923 年（建造这幢古老住宅的年份）减去（1850 + 1986）/2 = 1918（中值）得 7，除以年代范围的数值（1986 - 1850 + 1 = 137）生成一个比例值，然后乘以 2，得到的值是 0.0730。这个基本过程可应用于任何连续型特征，从而得到在 -1 和 1 之间的值。映射分类型特征的一种方式是在 -1 和 1 之间为每个类给出一个分数值。在本例数据中，惟一的分类变量是取暖类型 Heating_Type，因此，可以任意地映射 B 为 1，A 为 -1；假如有三个值，就可以分配一个为 -1，另外一个为 0，第三个为 1，尽管这种方法的缺点是：似乎三种取暖类型存在一个顺序关系，类型 -1 距类型 0 的距离显然比距类型 1 更近。第 17 章中进一步讨论了如何将分类变量转换成数值变量，而不会增加伪信息。

使用这些简单的技术，有可能映射先前（见表 7-2）展示的样板房记录的所有字段并训练网络。训练过程就是通过训练集来调整权重的迭代过程，有时把每次迭代称为一代。

网络被训练以后，必须在验证集上测试每一代的表现。神经网络的较早几代在验证集上的表现往往比最终网络（训练集被最优化）更好，这是由于过度适应（overfitting，已在第 3 章中讨论过）造成的，也是神经网络的强劲有力的结果。神经网络实际上是一个通用近似器的实例，也就是说，任何函数都可以用适当的复杂神经网络来提供相近的结果。神经网络和决策树都具有这样的特点，而线性和逻辑回归则不具备，因为它们假设基本函数具有特别的形状。

与其他的建模方法相比，神经网络仅仅能学会在训练集中存在的模式，导致过度适应。为了找到未知数据的最佳网络，训练过程记住每代期间计算出的一套权重，最终的网络来自于在验证集上工作最佳的那一代，而不是在训练集上运转最佳的一代。

当模型在验证集上的表现令人满意的时候，神经网络模型就已经为应用做好了准备。它已经从训练实例学习，并且了解如何从所有输入计算销售价格。模型读取住宅描述性信息，经过适当的映射后产生输出。有一个忠告，输出本身是一个介于 0 和 1 之间（对于逻辑激活

函数）或 - 1 和 1 之间（对于双曲正切）的数，它需要被再次映射回销售价格范围，举例来说，数值 0.75 可以乘以范围的大小（$147 000），然后加上该类的基数（$103 000），从而得到评价值 $213 250。

7.3 用于定向数据挖掘的神经网络

上面的实例展示了神经网络最普遍的应用：建立分类或预测模型。这一过程的步骤如下：

1）识别输入和输出特征；

2）转换输入和输出值，使其限定在一个小范围内（- 1 到 1）；

3）采用适当拓扑（布局）建立网络；

4）在一个训练样本的代表性集合上训练网络；

5）使用验证集选择使误差减到最小的权重集；

6）用测试集评估网络，观察网络执行情况；

7）应用网络产生的模型预测未知输入的结果。

幸运的是，现在的数据挖掘软件能自动地执行大部分的步骤。虽然不必非常熟悉内部工作的知识，但成功使用网络仍存在一些关键问题。和所有预言性建模工具一样，最重要的问题是选择正确的训练集；其次，应以一种合适的方法表达数据，使网络识别模式的能力最大化；第三，解释来自网络的结果；最后，要了解有关它们如何运行的一些特殊细节，如网络拓扑学和参数控制训练，这有助于构造运行良好的网络。

使用任何预言模型或分类模型的危险之一是，当模型衰老时，它的时效性较差——神经网络模型也不例外。对于评价实例，神经网络已经以训练集的内容为基础，获得了关于使它能够从住宅描述预测评估值的历史模式。谁也无法保证目前的行情与上周、上个月或 6 个月以前（当训练集被抽取出来时）相匹配。新住宅买卖每天都在进行，正在创造和回应在训练集中没有出现的市场购买力。利率的上升或下降，或通货膨胀的增加，都有可能快速改变评价值。受两个因素的影响，神经网络模型更加难以保持最新。首先，模型不容易以规则的形式表达它本身，因此，当它已经变得过期时，也不容易看出来；其次，当神经网络退化时，它们一般会微妙地、不明显地降低性能。简而言之，模型逐渐过期，但人们并不总是清楚什么时候应该更新它。

解决的办法是将近期的更多数据纳入神经网络。一种方法是将同样的神经网络返回到训练状态，并开始输送新数值。如果网络仅仅需要调整结果，比如当网络相当接近精确时，而你认为可以通过给它更多较近的例子来改善准确度的时候，这是一个好方法；另一种方法是通过把新的例子加入训练集（也许是删除旧的例子）重新开始训练整个网络，或许甚至采用不同的拓扑（后面将进一步讨论网络拓扑），当市场行情可能已经发生巨变，从原来的训练集发现的模式不再适用的时候，这是一个合适的方法。

在第 2 章中描述的数据挖掘良性循环促进了从数据挖掘活动测定结果。这些测定有助于了解给定的模型在多大程度上易于老化衰减，以及神经网络模型什么时候应该被重新训练。

警告： 神经网络至多像用于产生它的训练集一样好。模型是静态的，为了使它保持最新和有效，必须通过把更多近期的例子加入训练集并再次训练网络（或训练新的网络），以完成显式升级。

7.4 神经网络是什么

神经网络由基本单元构成，这些基本单元以简化的方式模仿自然界发现的生物神经元行为，不管这些神经元是组成人类的大脑还是青蛙的大脑。例如有人宣称，青蛙的视觉系统里面有一个单元，可以因为飞行运动而被激发，还有另外一个单元，对应于飞行物大小而被激发。这两个单元都与一个神经元相连接，当这两个输入组合值高的时候，这个神经元被激发。而这个神经元正是另外一个神经元的输入，从而引发青蛙舌部的探出行为。

其基本观点是，每个神经元（不管是青蛙的或计算机中的）都有许多的输入，神经元把这些输入结合在一起给出单一输出值。在大脑中，这些单元可能连接专门的神经。计算机中则比较简单，这些单元只是被简单地连到一起（如图 7-3 所示），来自某些单元的输出被当作其他单元的输入。图 7-3 所示是前馈神经网络（feed-forward neural network）的实例，这意味着有一个从输入到输出的单向流通过网络，在网络中没有循环。

前馈网络对于定向建模是最简单的和最有用的网络类型。关于前馈网络，有三个基本问题要明确：

- 单元到底是什么？它们是如何工作的？即，激活函数是什么？
- 单元如何被连在一起？即，网络拓扑是什么？
- 网络如何学会识别不同模式？即，反向传播是什么？更概括地说，网络是如何训练出来的？

对这些问题的回答提供了解基本神经网络的背景资料，这种了解可以指导我们利用这种强有力的数据挖掘技术来得到最好结果。

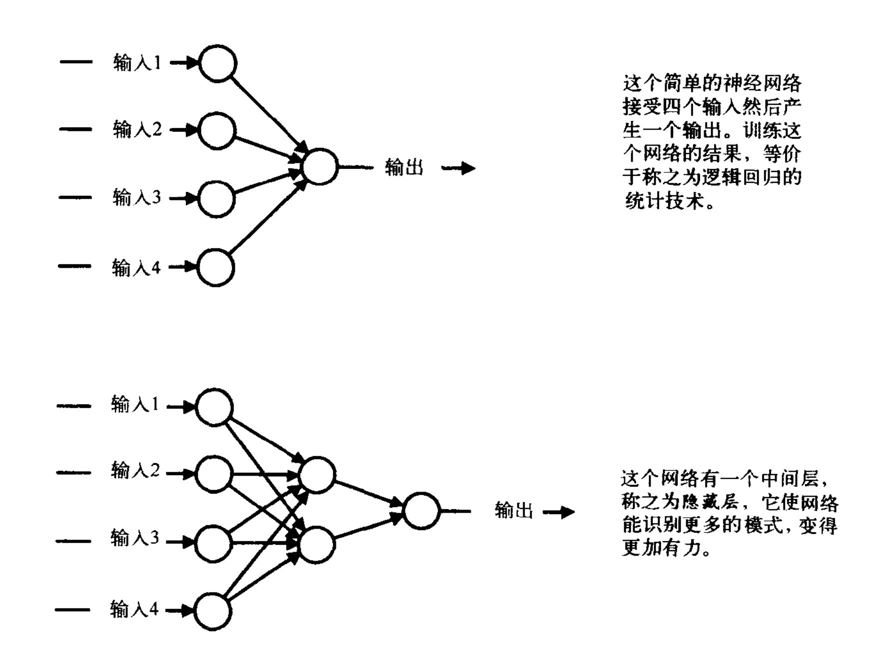


图 7-3 前馈神经网络从一端接受输入，然后把它们转变成输出

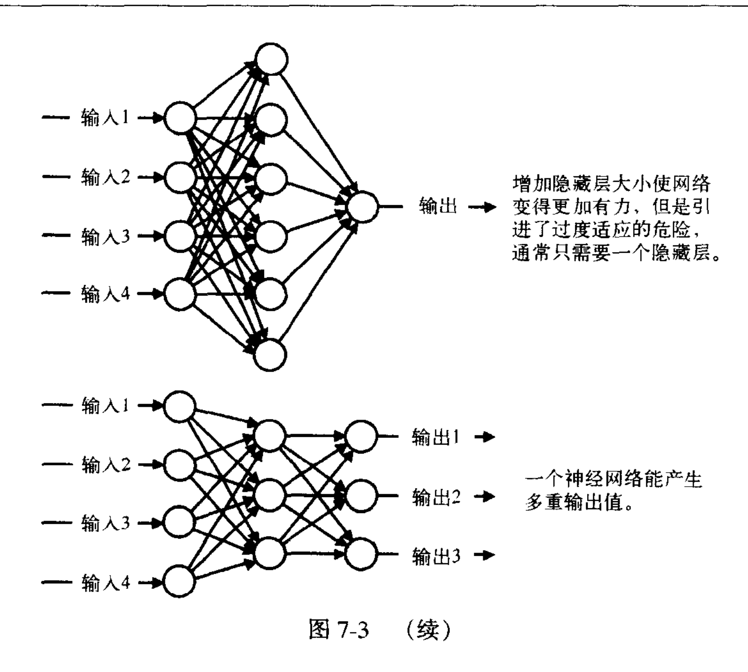


图 7-3 （续）

7.4.1 神经网络的单元是什么

图 7-4 显示人工神经元的重要特征。这个单元把几个输入结合到一起变成一个单一值，然后将它转换后产生输出，上述整个过程被称为*激活函数*。最通常的激活函数是以生物模型为基础，在组合的输入达到阈值以前，模型输出一直很低；当组合的输入达到阈值时，单元*被激活*，输出变得很高。

与它的生物学对应体相似，神经网络单元的特点是：当组合输入值处于某一中间范围时，输入的很小变化可以对输出产生较大的影响。相反地，当组合的输入远离中间范围时，大的输入值变化可能对输出影响甚微。这种特点，即小的变化有时很关键，但有时却不是这样，就是*非线性行为*的实例。神经网络的强劲有力和复杂性，都来自于它们的非线性行为，当然这也起因于组合神经元所使用的特定激活函数。

激活函数包括两个部分，第一部分是组合所有的输入成为单一值的*组合函数*（combination function）。如图 7-4 所示的那样，每个进入单元的输入有自己的权重，最通常的组合函数是加权和，即每个输入与它的权重相乘，然后把这些值求和。其他的组合函数有时是有用的，其中包括加权输入的最大值、最小值和值的逻辑“AND”或“OR”等。尽管在选择组合函数时有很大的灵活性，但使用标准的加权和在很多情况下工作良好，选择这个元素是神经网络的一个普通特点。它们的基本结构是相当灵活的，但是对应于最初的生物模型的那些默认值——如组合函数的加权和，在实际工作中运行良好。

激活函数的第二部分是*转换函数*，它因把组合函数的数值转化为单元的输出这一事实而得名。图 7-5 比较了三个典型的转换函数：Sigmoid（逻辑）、线性和双曲正切函数。转换函数所接纳的特定值不像一般函数那样重要。从我们的角度看，线性转换函数是最不令人感兴趣的。仅仅由带有线性转换函数的单元与权重组合函数总和构成的前馈神经网络只能做线性回归。Sigmoid 函数是 S 型函数，其中有两个最常用的神经网络函数——逻辑回归和双曲正切。它们之间的主要差异是其输出范围，逻辑回归介于 0 和 1 之间；双曲正切介于 −1 和 1 之间。

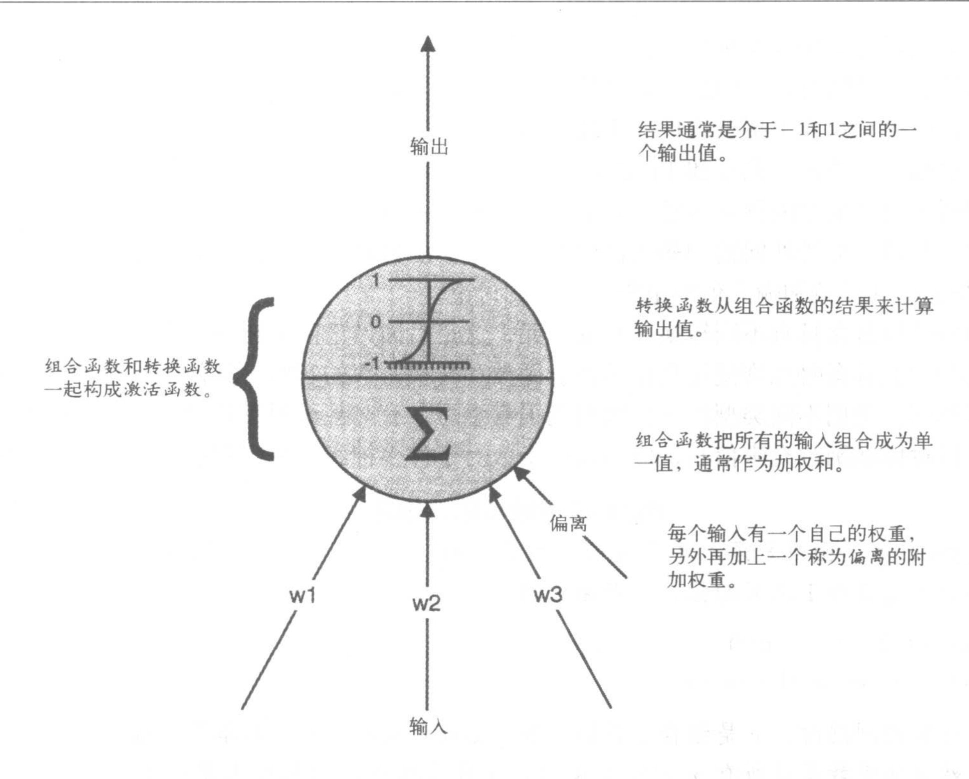


图 7-4　人工神经网络的单元是以生物神经元为基础建模的。单元的输出是其输入的一个非线性组合

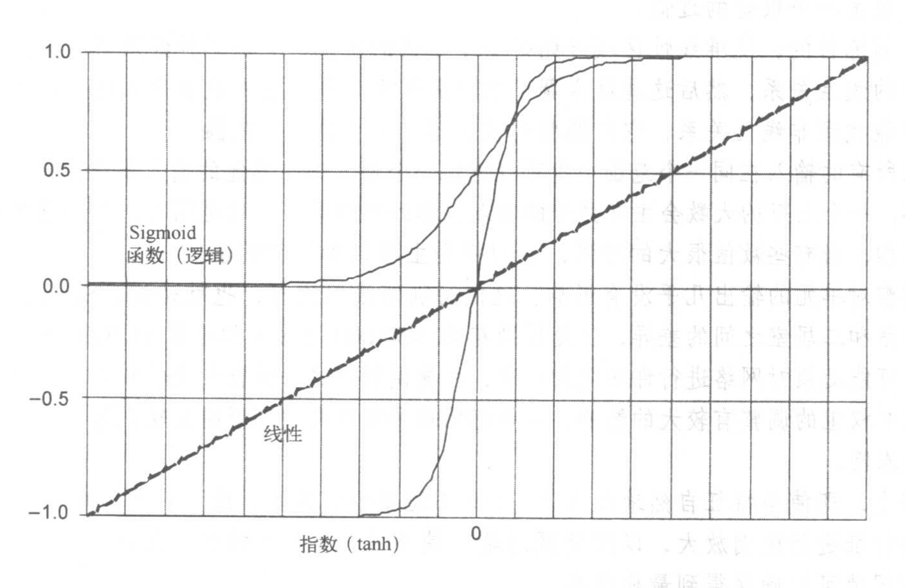


图 7-5　三个常用的转换函数是 Sigmoid 函数、线性函数和双曲正切函数

逻辑和双曲正切转换函数表现类似。尽管不是线性函数，但它们的表现还是引起了统计学家的注意。当所有输入的加权和接近 0 时，这些函数近似于线性函数。统计学家很喜欢线性系统，也很喜欢几乎接近线性的系统。如加权和的量级变得更大时，这些转换函数逐渐饱和（在逻辑回归情况下到 0 和 1；在双曲正切情况下到 -1 和 1)。这种行为符合从输入的线性模型到非线性模型的逐渐过渡。简而言之，神经网络有能力对三种类型的问题做好建模工作：线性问题、拟线性问题和非线性问题。在激活函数和输入值范围之间也有一个关系，见后面“Sigmoid 函数和输入值的范围”部分的讨论。

网络可以包含具有不同转换函数的单元，这是后面讨论网络拓扑时会再次讨论的一个主题。复杂的工具有时允许使用其他的组合函数和转换函数的实验。其他函数与标准函数的行为显著不同，使用不同类型激活函数可能很有意思，有时甚至很有帮助，但如果你不想找麻烦，可以对标准函数充满信心，因为这些函数对于许多神经网络应用已被证明是成功的。

Sigmoid 函数和输入值的范围

Sigmoid 激活函数是落入某界线内的 S 形曲线。比如，对于所有求和函数的可能输出，逻辑函数产生 0 和 1 之间的值，而双曲正切产生 -1 和 1 之间的值。这些函数的公式是：

$$\text{logistic}(x) = 1/(1+e^{-x})$$

$$\tanh(x) = (e^{x}-e^{-x})/(e^{x}+e^{-x})$$

用于神经网络时，x 是组合函数的结果，比较典型的是取进入单元的输入的加权和。

既然这些函数是对所有 x 的值定义的，为什么推荐网络的输入是一个小的范围，通常从 -1 到 1？理由与这些函数在 0 附近的表现有关。在这个范围中，它们的行为几近线性，即 x 的微小变化导致输出的微小变化；x 变化一半导致输出变化大约一半。这个关系不是精确的，但是一个很好的近似。

出于训练目的，从准线性区域开始是一个不错的主意。当训练神经网络时，结点可能找到数据中的线性关系，然后这些结点调整它们的权重，使产生的数值落入这个线性范围；其他结点可能找到非线性关系，它们调整的权重落在一个较大的范围。

要求所有的输入在同一个范围中也可以避免一个输入集在其他的输入集中占有优势，例如住宅价格，一个上万的大数会主导另外的输入，如卧室的数目。这是因为，组合函数毕竟是输入的加权和，当有些数值很大的时候，它们将会主导该加权和的值。当 x 很大时，输入权重的微小调整对单元的输出几乎没有影响，这使得训练难以进行，也就是说，Sigmoid 函数可以利用一居室和二居室之间的差异，但是区别花费 $50 000 的住宅和花费 $1 000 000的住宅可能很困难，可能必须对网络进行许多代的训练，才能调整与这个特征相关的权重。保持相对小的输入可以对权重的调整有较大的影响，对训练的这种帮助是我们坚持把输入限定在一个小范围的最重要原因。

事实上，即使当特征自然地落入比 -1 到 1 还要小的范围，比如 0.5 到 0.75，我们也希望把这些特征进行比例放大，以便使网络输入使用从 -1 到 1 的整个范围。使用从 -1 到 1 的整个范围值可以确保得到最佳结果。

虽然推荐的输入范围是从 -1 到 1，但这应该视为一种方针，而不是严格的准则。举例来说，标准化变量——减去均值后除以标准差（standard deviation）——就是一个常用的变量转换，这为神经网络产生足够小的有用数值。

7.4.2 前馈神经网络

前馈神经网络从输入值来计算输出值，如图 7-6 所示。这个网络拓扑（或结构）是用作预测和分类的典型网络。单元被编入三个层，在左边的层与输入相连接，被称为*输入层*，输入层的每个单元只与一个源字段相连，通常映射在 - 1 到 1 的范围。在这个实例中，输入层实际上没起任何作用，每个输入单元只是将输入值复制变成输出值。如果情况是这样，为什么要不厌其烦地在这里提及呢？因为它是神经网络词汇表的重要组成部分，在实际术语中，输入层代表将值映射到合理范围的过程。正是由于这个原因，有必要把它们包括进来，因为它们暗示了成功使用神经网络的一个很重要的方面。

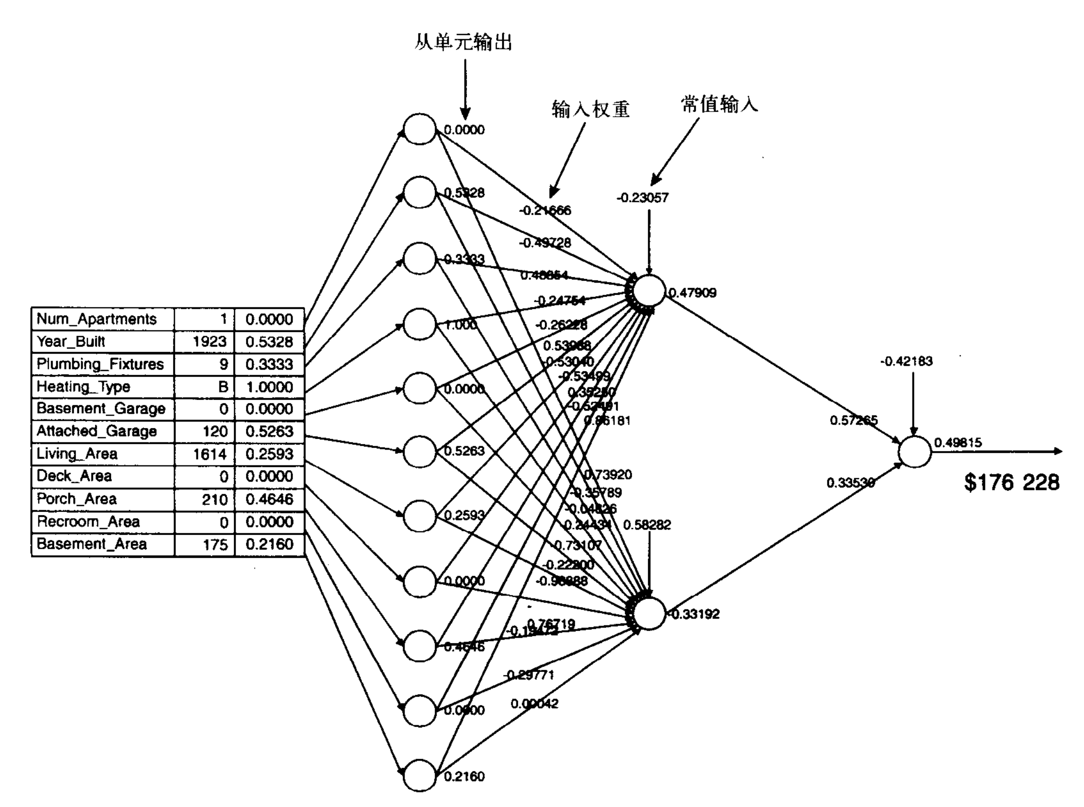


图 7-6　此处所显示的房地产训练实例，将输入提供到前馈神经网络，表明网络充满看似无意义的权重

下一层被称为*隐藏层*，因为它既不与网络输入相连，也不与网络输出相连，隐藏层的每个单元通常与输入层的所有单元相连接。由于这个网络包含许多标准单元，隐藏层的众多单元把每一个输入值乘以对应的权重，然后将这些值求和，最后运用转换函数计算各自的输出值。神经网络允许有任意数目的隐藏层，但通常来说，有一个隐藏层就足够了。这个层越宽泛（即包含较多的单元），网络识别出模式的能力就越高。然而这个更高的能力也存在缺陷，因为神经网络可能记住在训练实例中的某一种模式。*我们希望网络能够从训练集中得到总结，而不是记住它*，为达到这个目的，隐藏层不应该太宽。

注意图 7-6 中的那些单元，每个单元都有来自顶端的附加输入。这是固定输入，有时称

为偏离，且总是被设定成 1。像其他输入一样，它有权重且包含在组合函数中，偏离充当整体偏移以便帮助网络较好地理解模式。训练阶段调整固定输入的权重的方法与网络对其他权重的调整方法相同。

右边的最后一个单元就是输出层，与神经网络的输出相连接，它也与隐藏层中的所有单元相连接。多数情况下，神经网络是用来计算单一值，因此在输出层中只有一个单元和一个值。我们必须把这个数值映射回来以便理解输出结果。对于图 7-6 的网络，我们必须把 0.49815 这个数值转换回到一个在 \$103 000 和 \$250 000 之间的值，它对应的是 \$176 228，实际上非常接近实际价值 \$171 000。在有些执行过程中，输出层使用简单的线性转换函数，因而输出是输入的加权线性组合，这就去掉了将输出进行映射的必要。

输出层可以有一个以上的单元，举例来说，一家连锁百货公司想要预测客户将会购买不同部门产品的可能性，如女士服装、家具和娱乐产品等，以便利用这种信息来策划促销活动，以及进行直接目标邮寄。

为了做出这种预测，可以建立如图 7-7 中所示的神经网络。这个网络有三个输出，每个部门对应一个输出，输出结果是在输入中描述的客户从相关部门再次购买的倾向。

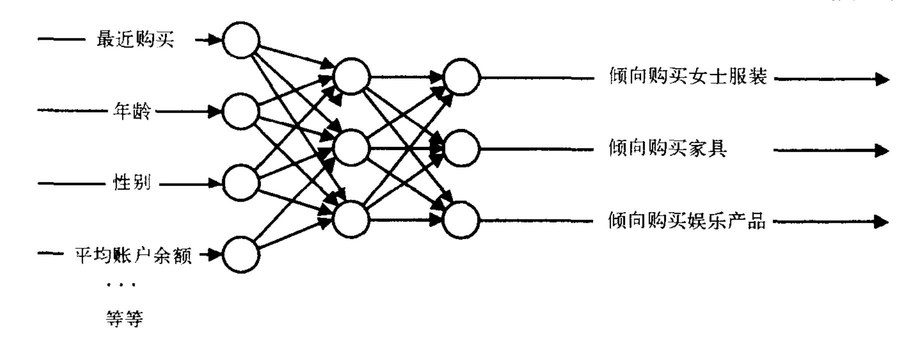


图 7-7 该网络有多个输出，可用来预测百货公司客户下次将会在哪个部门购买

当把一个客户的输入送进网络以后，神经网络计算出三个值，在得到这些输出结果以后，百货公司如何决定向客户提供正确的一种或者多种促销呢？当处理若干模型的输出结果时，一些常用的方法是：

- 记下与最大值输出对应的部门；
- 记下与前三名的输出数值对应的部门；
- 记下与超过某个阈值的输出对应的所有部门；
- 记下那些大小为最大值单元某个百分比的单元对应的所有部门。

所有的这些可能性都能起到很好的作用，每一种各有不同的强项和弱点，没有一个总是适合任何情况的正确答案。实际上，需要在测试集上实验这些可能性，以便决定在特定的情形下哪一种最好。

前馈神经网络拓扑有其他的变形，输入层有时被直接连到输出层。在这种情况下，网络有两个部分：这些直接的连接像一个标准回归一样（线性或逻辑的，取决于输出层的激活函数），对于构建更标准的统计模型是有益的，隐藏层随后对统计模型进行调整。

7.4.3 神经网络如何使用反向传播学习

训练神经网络的过程就是设定连接网络所有单元之间的边的最佳权重。目标是使用训练

集来计算权重，对于训练集中尽可能多的实例，使得网络的输出尽可能接近期望结果。虽然反向传播不再是调整权重的优先方法，但是它提供关于训练如何运作的信息，并且它是训练前馈网络的原始方法。反向传播的核心是以下三个步骤：

1）网络得到训练实例，使用网络现有的权重，计算一个或多个输出；

2）然后，反向传播通过计算结果和预期结果（实际结果）之差来计算误差；

3）通过网络反馈误差，并且调整权重将误差减至最小——由于误差经过网络被返回，所以得名“反向传播”。

反向传播算法通过比较每个训练实例产生的值和实际值，测定网络的总体误差，然后调整输出层权重，减小但不是消除误差。然而算法到此并未结束，它把误差责任归咎于早期的网络结点，调整连接那些结点的权重，进一步减小总体误差。分担责任的特定机制并不重要，它足以说明反向传播使用复杂的数学过程，需要取激活函数的偏导数。

给定误差，单元如何调整它的权重？它估算变更每个输入的权重是否会增加或减小误差。于是，单元调整权重，减小但不是消除误差。训练集中每个实例的调整慢慢影响权重，以期获得最佳值。要记住，网络的目标是总结和识别输入的模式，而不是记住训练集，因而调整权重要像悠闲的散步，而不是疯狂疾驰的短跑。在足够多的代以内遇到足够多的训练实例之后，网络权重不再有显著改变，误差也不再减小，这个点就是训练终止点，此时网络已经学会在输入中识别模式。

调整权重的技术被称为*通用的 δ 规则*，有两个重要参数与应用通用的 δ 规则相关联。第一是*动量*（momentum），它涉及在每个单元内权重向某个“方向”改变的趋向，即每个权重记住自己是否已经变得更大或更小，并且动量设法使它在相同的方向保持发展。带有高动量的网络对于希望翻转权重的新的训练实例响应缓慢；如果动量低的话，允许权重更自由地振荡。

训练的最优化

虽然反向传播是训练网络的第一个实用算法，但它效率低。训练的目标是要找到将训练集和（或）验证集上的误差减到最少的权重集。这一类型的问题属于最优化问题，有几种不同的方法可以做到这一点。

值得注意的是，这是一个难题。首先，网络有许多权重，因此，需要考虑许多种不同权重的可能性。对于有 28 个权重的网络（假设在隐藏层有 7 个输入和 3 个隐藏结点），如果尝试每个权重只有两个值的组合，需要测试 2^{28}次组合，即超过 250 000 000 次组合，因而对于每个权重尝试所有 10 个值的组合代价会极其昂贵。

第二个问题是对称性。一般来说，没有单一的最优值。事实上，对于隐藏层有一个以上单元的神经网络，总是有多个最优值（optimal），因为在一个隐藏单元上的权重可能完全与另一个单元上的权重交织在一起。有多个最优值的问题使得发现最优结果变得复杂。

发现最优值的一种方法被称为爬山法。从一个随机权重集开始，然后在每个方向上采取单独的一步，对每个权重做一点小的改变，选择出能够最好地减少误差的任何一小步，并重复这个过程。这就好像一步一步向山上爬，从而发现山的某处是最高点一样，但在许多情况下，你可能结束于小山包的顶峰而不是高山的顶峰。

爬山的另一种方式是从大步伐开始，然后逐渐地减小步幅（巨人 Jolly Green 可能会比一只蚂蚁更轻松地到达最近的山峰）。一个相关的算法被称为模拟退火（simulated annealing），在登山过程中引入一点随机性。随机性以物理理学领域为基础，与晶体如何由液态冷

却形成固态相关（水晶的形成是物理界中一个最佳实例）。模拟退火和爬山两者都需要许多次迭代，并且这些反复的迭代从计算角度看也是昂贵的，因为需要在整个训练集上运行网络，并且对每一步进行一次又一次的重复计算。

用于训练的较好算法是共轭梯度（conjugate gradient）算法。这个算法测试几个不同的权重集，然后推测最适当的位置，这需要使用多维空间几何学的一些理念。每个权重集被视为多维空间的一点，在尝试多个不同的集合后，算法决定匹配这些点的一条多维空间抛物线。抛物线是 U 形曲线，有惟一的最小值（或最大值）。然后，共轭梯度在这个区域中利用新的权重集继续进行训练。这个过程仍然需要反复；然而，与反向传播或各种爬山方法相比，共轭梯度能够更快地产生较好的值，共轭梯度（或它的一些变体）是多数数据挖掘工具中训练神经网络的首选方法。

学习速率控制权重变化的快慢。当训练网络时，改变学习速率的最佳方法是从大的开始，然后在网络训练过程中逐渐减小。最初，权重是随机的，因此在最佳权重附近出现大的振荡是非常有用的。然而，当网络逐渐靠近最佳解决方案时，学习速率应该减小，以便网络对最佳权重进行精细地调整。

研究人员已经为训练神经网络发明了数以百计的变换方法（参见前面“训练的最优化”部分），其中每个方法都有其优缺点，但它们都是在寻找训练网络很快达到最优方案的技术。有些神经网络包提供多重训练方法，允许使用者通过实验获得最佳方案。

采用任何训练技术的危险之一是陷入被称为“局部最优”的某个状态。当网络为训练集产生好的结果，而且调整权重不再促进网络的性能时，就可能发生这件事情。然而，还有一些权重的其他组合与网络的这些权重显著不同，却产生非常好的解决方案。这就类似设法爬到山顶的运动，每回都尝试选择最陡峭的道路，却发现仅仅攀登到附近的小山顶上，因为在发现局部最佳方案和整体最佳方案之间有一个张力（tension）。控制学习速率和动量有助于找到最佳解决办法。

7.4.4 前馈网络和反向传播网络的启发

即使有尖端的神经网络包，要得到神经网络的最好输出结果仍需一些努力。本部分涵盖了搭建网络以获得好方案的一些探索。

或许，最需要确定的是隐藏层的单元数目。单元越多，网络识别的模式也越多，这会使我们力争建立更大的隐藏层。然而，这样做有一个缺陷，网络可能最终会记住它，而不是对训练集进行总结。在这种情况下，更多并不意味着更好。幸运的是，你可以发现什么时候网络被过分训练：如果网络在训练集上运行得很好，但在验证集上运行得相当糟，这说明它已经记住训练集。

隐藏层应该是多大呢？没人知道真实的答案。这取决于数据、要寻找的模式以及网络的类型。因为过度适应是使用客户数据网络的主要问题，通常，隐藏层数目要少于输入数目。对许多问题来说，一个好的开端是在隐藏层实验一个、两个和三个结点，这是可行的，尤其是现在训练神经网络仅需要几秒或几分，而不是数小时的时间。如果增加较多结点改善网络的表现，那么隐藏层更大可能更好。当网络已经被过分训练的时候，需要减少层的大小；如果它不足够准确，就增加层的大小。然而，使用网络进行分类时，最好对每个类从一个隐藏结点开始训练。

另一个需要确定的是训练集的大小。训练集一定要足够大，足以覆盖每个特征的有效输入范围，除此之外，对网络的每个权重，需要多个训练实例。对于有 s 个输入单元，h 个隐藏单元，1 个输出的网络，在网络中有 $h*(s+1)+h+1$ 个权重（每个隐藏层结点有一个权重对应于输入层的连接，一个作为偏离的附加权重，然后是一个与输出层的连接和它的偏离）。例如，如果在隐藏层网络中有 15 个输入特征和 10 个单元，于是在网络中有 171 个权重。每个权重至少应该有 30 个实例，但最好至少有 100 个实例，所以对于这个例子，训练集至少应该有 17 100 个实例。

最后，要使用反向传播训练算法从网络中获得好的解决方案，学习速率和动量参数是很重要的（最好使用共轭梯度或类似方法）。最初，学习速率应该设为高，以便对权重进行大调整。然后，为了精细调整网络，应该降低学习速率。动量参数允许网络以更快速度向解决方案移动，以防止围绕不太有用的权重振荡。

7.5 选择训练集

训练集由预言或分类值已知的一些记录组成。对于所有的数据挖掘建模，选择好的训练集至关重要，即使不考虑参与创造训练集的其他行为，糟糕的训练集注定了网络的命运。幸运的是，在选择好的训练集时只需要考虑几件事情。

7.5.1 覆盖所有特征值

在所有需要考虑的事情中最重要的是，训练集需要覆盖网络可能遇到的所有特征取值的完整范围，也包括输出。在房地产评估实例中，这意味着包括便宜的住宅和昂贵的住宅、大住宅和小住宅，以及带有车库的住宅和不带车库的住宅，等等。总的来说，对于每个分类特征取值和遍布整个有序离散或连续特征取值范围的值，训练集中都应该有相应的实例。

不管这些特征是否确实被作为输入输送到网络，覆盖所有特征值都是正确的。例如，在神经网络中，占地面积大小不可能作为输入变量。然而，训练集仍然应该具有所有不同占地面积大小的实例。一个在较小占地面积上训练的网络（有些可能是价位定低了，而有些是价位定高了）不可能对豪华庄园式住宅做出出色的工作。

7.5.2 特征数目

输入特征的数目会以两种方式影响神经网络。首先，输入网络的特征越多，网络就需要越大，这样就增加了过度适应的危险和增大了训练集的大小。其次，特征越多，将网络收敛到一组权重所需的时间越长。而且对于太多的特征，权重不大可能达到最佳。

这一变量的选择问题是统计学家关心的普遍话题。实际上，我们发现决策树（在第 6 章中已讨论过）为选择最佳变量提供一个好的方法。图 7-8 显示的是 SAS Enterprise Miner 的一个良好特征。通过将神经网络结点连接到决策树结点上，神经网络可以只使用决策树选择出的变量。

另外一种方法是使用直觉，从少数变元开始是明智的。通过实验尝试其他变量，观察哪些变量能改善模型。在许多情况下，计算能够代表商业问题的特殊方面的那些新变量是非常有用的，如在房地产实例中，我们可能从占地面积大小减去住宅的大小来计算院子的大小。

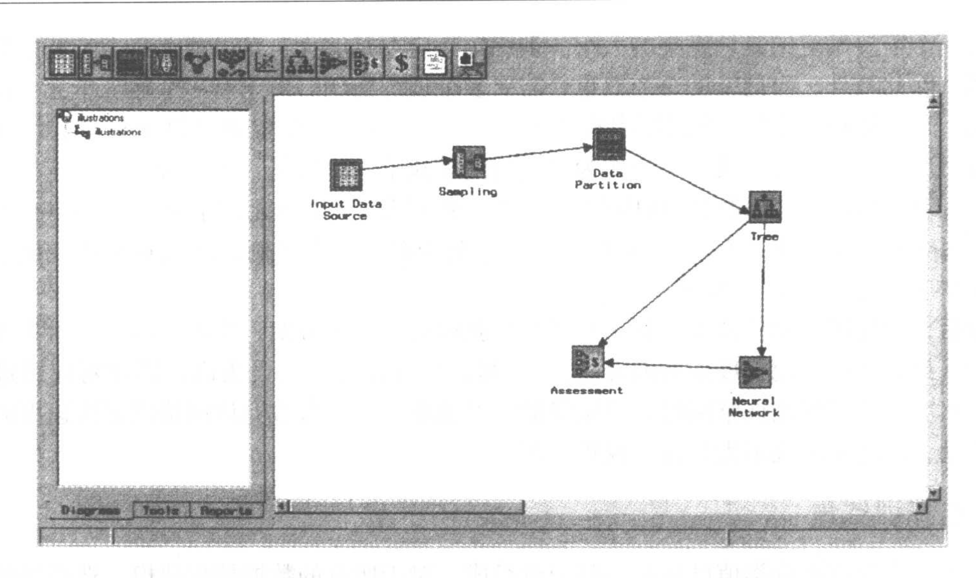


图 7-8 SAS Enterprise Miner 提供了一个简单的选择神经网络变量的机制——只要将神经网络结点连接到决策树结点就可以了

7.5.3 训练集的大小

网络中的特征越多，为覆盖数据模式的有效范围需要的训练实例就越多。不幸的是，没有简单的规则表示特征数目和训练集大小之间的关系。然而，一般情况下，最少需要有几百个实例支撑每个具有有效范围的特征，当然数千个也不是不切实际的。作者曾经处理过只有 6 个或 7 个输入的神经网络，但是，它的训练集包含了数十万个列。

当训练集不够大的时候，神经网络倾向于过度适应数据。当训练实例比网络权重还少时，过度适应肯定会发生。这将带来一个问题，就是网络在训练集上运转非常好，但是在未见数据上则遭遇惊人的失败。

当然，非常大的训练集也有不利的一面，它需要花费较长的时间训练神经网络。在给定的时间内，为了得到较好的模型，可以使用较少输入特征和较小的训练集，并且用不同的特征组合和网络拓扑进行实验，而不是使用最大的训练集（使用最大的训练集就可能没有时间做实验）。

7.5.4 输出数目

在多数训练实例中，通常的情况是：参与进来的输入比给出的输出多得多，因此好的输入覆盖范围导致好的输出覆盖范围。然而，对于网络的所有可能的输出值，有许多实例是很重要的；除此之外，对每个可能的输出其训练实例数目应该在数量上大体相当。当决定什么可以作为训练集的时候，这可能是很关键的。

例如，如果神经网络被用于探测罕见但很重要的事件——如柴油机故障率、信用卡的欺诈使用或者谁将响应家庭债券信用度的计划——那么训练集必须有这些罕见事件的足够多实例。有效数据中的随机样本可能并不够，因为普通的例子会淹没罕见实例。为了解决这个问题，训练集需要通过过度采样罕见案例从而达到平衡。对于这类问题，由 10 000 个“好”

的实例和10 000个"坏的"实例组成的训练集，比随机挑选100 000个好例子和1 000个坏例子组成的训练集效果会更好。毕竟，当使用随机抽取样本的训练集时，神经网络可能会忽略输入，直接标示"好的"，并且接近99%的时间都正确。对于"训练集越大就越好"这一普遍规则而言，这是一个例外。

提示： 神经网络训练集必须足够大，以便能覆盖全部特征的取值。对于每个输入特征，就算不需要数百或数千，也至少需要十几个实例。对于网络的输出，你要确信值会平均分布。这是一个较少的训练集实例实际改善结果的案例，当你想要训练它识别"坏"实例时，不会被网络中的"好"例子所淹没。训练集的大小也受运行模型所使用机器能力的影响，当训练集很大的时候，神经网络需要较长的时间加以训练，这些时间也许用来改善对不同的特征、输入的映射函数和网络参数下的实验会更好。

7.6 准备数据

准备输入数据经常是使用神经网络最复杂的内容。这种复杂性一部分体现在数据挖掘试图选择正确数据和正确实例，另外一部分体现在需要把每个域映射到适当的范围——记住，使用有限的输入范围有助于帮助网络较好地识别模式，有一些神经网络软件包通过使用友好的图形模式界面来提供方便的转换。因为进入网络的数据格式对神经网络执行性能有很大的影响，我们来看看映射数据的普遍方法。第17章包含关于数据准备的另外一些材料。

7.6.1 具有连续数值的特征

有些特征具有连续数值，通常位于已知的最小界限和最大界限之间。这类特征的例子是：

- 美元数额（销售价格、月结余额、周销售量、收入等）
- 平均数（月平均结余额、销售量平均数等）
- 比率（债务收入比、价格收入比等）
- 物理度量（生活圈、温度等）

房地产评估实例显示了一个处理连续特征的好方法。当这些特征落入被预先定义的最小值（min）和最大值（max）之间时，数值能被按比例缩放到一个合理的范围中，比如应用如下的计算：

```
mapped_value = 2 * (original_value - min) / (max - min + 1) - 1
```

这个转换（减去最小值，除以范围，乘以2再减1）产生-1到1之间的一个数值（mapped_value），该值遵循与初值（original_value）同样的分布。在许多案例中，该转换工作良好，但是需要一些额外的考虑。

首先要考虑的是，训练集变量的取值范围可能不同于被评价数据的范围。当然，可以通过确保变量值在训练集中有代表，以尽量避免这种情况。然而，这种理想的情况并不总是可行的。有人可能在附近建造生活空间是5 000平方英尺的新住宅，这可能导致房地产评估神经网络变为无用。解决这个问题有多种方法：

- 制定较大范围的计划。在训练集中，住宅的生活空间范围被设置为714平方英尺到4185平方英尺。但我们可以不使用这些值作为最小值和最大值，而是允许有一定弹

性，比如设为500到5000。

- 放弃超出范围的值。在训练集中，一旦启用超出数值范围的外围数据，我们对结果的信心就更少。记住，仅将网络用于预定义输入值的范围，当为控制制造过程使用神经网络的时候，这一点尤其重要。不正确的粗糙结果会带来灾难性后果。
- 低于最小值的值用最小值代替，高于最大值的值用最大值代替。因此，所有超过4 000平方英尺的住宅视为4 000。这在许多情况下是有效的。然而，住宅价格与生活空间大小密切相关，所以具有超过最大住宅20%生活空间的住宅（所有其他的是相同的）大约会多花费20%的钱。在其他情况下，限定有关的数值效果很好。
- 把最小值映射到－0.9，把最大值映射到0.9，从而代替－1和1。
- 或者，最可能的是无需担忧。大多数值靠近0是很重要的，几个例外或许将不会造成重大的冲击。

图7-9展示由连续特征带来的另外一个问题：数值非对称分布。在这一数据分布中，几乎所有的收入都低于 \$100 000，但范围是从 \$10 000 到 \$1 000 000。按照建议比例，映射收入数值 \$30 000 为－0.96，收入\$65 000对应－0.89，几乎没有差别，但对销售应用，这个收入差别是很明显的。另一方面，\$250 000 和 \$800 000 分别对应－0.51 和＋0.60，有很大的差异，尽管这个收入差异可能并不十分明显。收入向低端高度倾斜，这会让神经网络难于利用收入字段，非对称分布能阻碍网络有效地使用重要的字段。非对称分布影响神经网络但不影响决策树，因为神经网络实际上使用数值进行计算，而决策树只使用数值的排序（等级）。

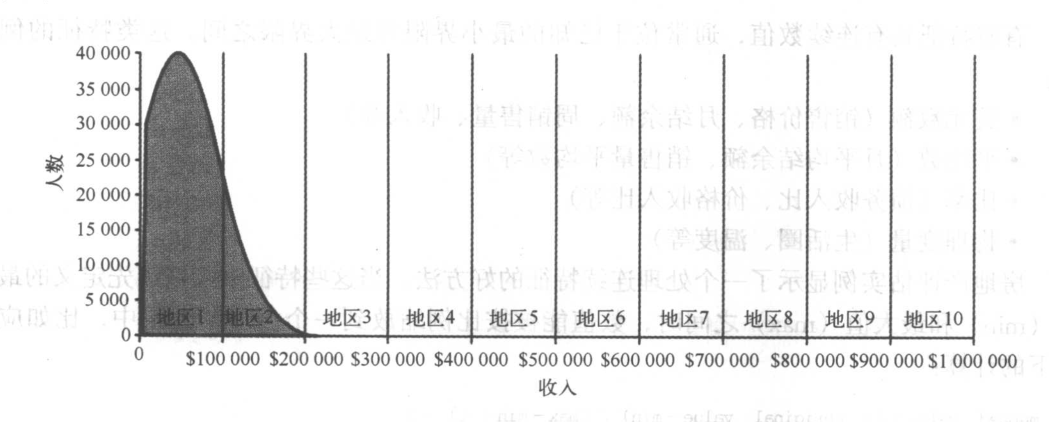


图7-9　家庭收入提供了非对称分布的实例。几乎所有值落在开始的10%范围内（收入小于 \$100 000）

有几种方法可以解决这个问题。最普通的做法是拆分收入特征，把它分成不同范围，这被称为将字段离散化（discretizing）或分箱（binning）。图7-9表示了将收入分解成10个等值的范围，但是这毫无用处，事实上，所有值落入前两个范围。五等分值提供了较好的选择范围：

- \$10 000～\$17 999　　很低（－1.0）
- \$18 000～\$31 999　　低（－0.5）
- \$32 000～\$63 999　　中等（0.0）

- \$64 000～\$99 999　　高（+0.5）
- \$100 000 及以上　　很高（+1.0）

这个转换存在信息丢失，一个收入 \$65 000 的家庭现在看起来完全像一个收入 \$98 000 的家庭。但另一方面，十分巨大的值不会给神经网络造成混乱。

当然还有其他的方法。例如，取对数是处理宽范围数值的好办法。还有一个方法是将变量归一化（减掉均值和除以标准差）。归一化的值时常在 -2 和 +2 之间（即对于多数变量，几乎所有数值落入均值的两个标准差之间）。对神经网络来说，归一化变量通常是一个好方法。然而，因为大的异常值使标准差变大，所以一定要小心地使用。因此，当有大的异常值时，许多归一化数值将会落入很小的范围，使网络很难区分它们。

7.6.2 具有有序、离散（整数）数值的特征

连续特征能归档成有序的离散数值。具有有序离散数值特征的其他例子包括：

- 计数（孩子的数目、购买物品的数量、购买后的月数等）
- 年龄
- 已排序的类（低、中、高）

像连续型特征一样，它们也有最大值和最小值。例如，通常年龄范围大约从 0 到 100，但是精确的范围可能依赖于使用的数据。孩子的数目可能从 0 到 4，超过 4 被看做 4。准备此类字段很简单：首先，计算出不同值的个数，并且在某一范围（如从 0 到 1），给每个值分配一个比例分数值，例如，如果有 5 个截然不同的值，就把它们映射为 0，0.25，0.50，0.75 和 1，如图 7-10 所示。注意，把数值映射到这样的单元区间保持了它们的排序，这是该方法的一个重要方面，意味着信息没有丢失。

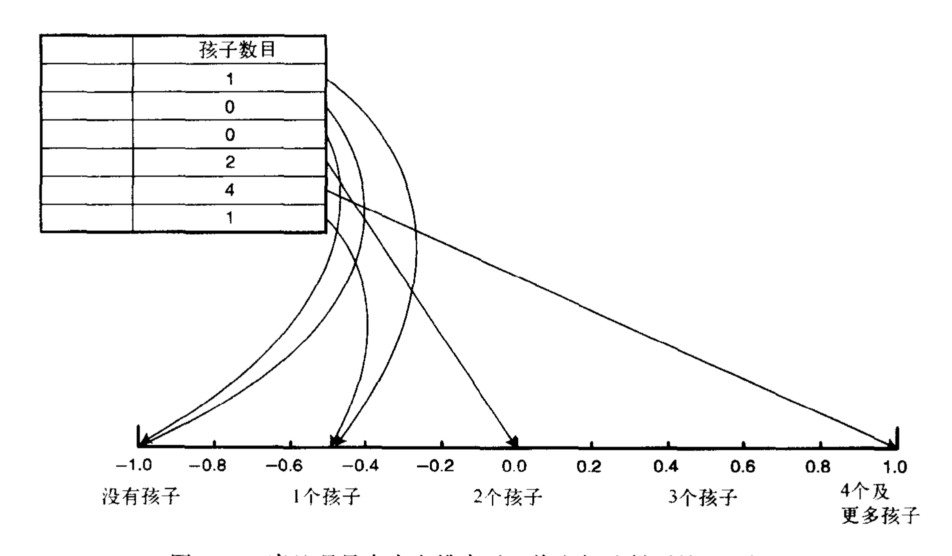


图 7-10　当编码具有内在排序时，将它们映射到单元区间上

也可以将范围划分为不等的部分，一个被称为温度计编码（thermometer code）的实例是：

0→0000 = 0/16 = 0.0000

1→1000 = 8/16 = 0.5000

2→1100 = 12/16 = 0.7500

3→1110 = 14/16 = 0.8750

这个案例名字的起因是，1的序列从一边开始，上升到某个值，就像温度计内的汞。然后这个序列被解释成以二进制表示的十进制数。温度计编码对学术评级和债券评估是有用的技术，因为其中一端标度的差异不如另一端明显。

例如，对于许多销售应用，没有孩子与有1个孩子有质的差别。然而，有3个孩子和有4个孩子的差别就可以忽略。使用温度计编码，孩子变量的数目可能映射如下：0（对应于0个孩子），0.5（对应于1个孩子），0.75（对应于2个孩子），0.875（对应于3个孩子），等等。对于分类变量，经常更容易将映射值保持在0到1的范围内，这是合理的。不过也可以将范围扩展到-1到1，只需要把值乘以2减1即可。

温度计编码是将以前信息包括进编码系统的一种方法，保持特定的编码值且紧密靠拢，原因是你感觉这些编码数值应该紧密靠拢在一起。这类知识能改善神经网络的结果——不要让它发现你已经知道的信息。灵活地把数值映射到单元区间上，以便彼此接近的编码与你对它们应该接近程度的直觉相匹配。

7.6.3 具有分类数值的特征

分类特征是数值的无序列表（unordered list）。它们不同于有序列表，因为其中没有需要保持的排序且不适宜引入顺序。通常有许多数据属于分类数值的实例，例如：

- 性别、婚姻状况
- 状态代码
- 产品代码
- 邮政编码

在美国，虽然邮政编码看起来是一组数字，但它们实际上代表的是离散的地理区域，并且代码本身提供的地理信息很少。没有理由认为10014比94117更接近02116，尽管从数字上看非常靠近。数字代表的只是与地理区域相关联的离散的名字。

处理分类特征基本上有三种不同方法。第一种方法是，视编码为离散的有序值，使用前面讨论的方法进行映射。不幸的是，神经网络不能理解这是无序的编码。因此，婚姻状况的五个代码（“单身”、“离婚”、“已婚”、“守寡”和“未知”）会分别被映射到-1.0，-0.5，0.0，+0.5，+1.0。

从神经网络角度看来，“单身”和“未知”两者相差甚远，然而“离婚”与“已婚”相当接近。对于有些输入字段，这种隐含的排序可能不会产生多大的效应。而在有些案例中，数值间彼此有某种相关性，隐含的排序混淆了网络。

警告：当使用神经网络中的分类变量时，把变量映射为数值要很小心。映射引进了变量的排序，神经网络会考虑到这种排序，即使排序本身没有任何意义。

第二种处理分类特征的方法是把类分解成标志，每个类有一个标志。假定性别有三个不同的值（男性、女性和未知），表7-3显示三个标志如何使用一个被称为N中选1的方法来对这些数值进行编码。可以通过除去性别未知标志从而减少标志的数目，这种方法称为$N-1$中选1的编码。

表 7-3 使用 N 中选 1 编码和 N－1 中选 1 编码处理性别

性别	N 编码			N－1 编码	
	男性标志	女性标志	性别未知标志	男性标志	女性标志
男性	+1.0	-1.0	-1.0	+1.0	-1.0
女性	-1.0	+1.0	-1.0	-1.0	+1.0
未知	-1.0	-1.0	+1.0	-1.0	-1.0

为什么要这么做呢？现在，我们已经放大了输入变量的数目，这对神经网络来说通常是一件坏事情，不过，这些编码方案是惟一能够除去数值隐含排序的方法。

第三种方法是，用有关编码的数字型数据替换编码本身。不在模型中包括邮政编码，而是包括不同的人口调查字段，如中值收入或有孩子的家庭比例。另外一个可能性是，把在分类变量层次上汇总出的历史信息包括进来。一个例子是用于预测流失的模型中，包含按照邮政编码给出的历史流失率。

提示： 当在神经网络中使用分类变量时，尽量用描绘它们的数字变量替换之，例如在人口调查区域的平均收入，在一个邮政编码（穿透度）内的客户比例、手机客户的历史流失率或定价计划的基本成本。

7.6.4 其他类型的特征

有些输入特征可能不能直接归入这三个类。对于复杂的特征，需要提取有意义的信息和使用上述技巧之一描述结果。记住，神经网络的输入通常介于－1 和 1 之间。

日期是需要以特别方式处理的一个很好的数据实例。任何日期或时间都能相对于某个固定点以天数或秒数等数字描述，它们可以被映射并直接输入到网络中。然而，如果日期是用于转账，那么每周的第几天和每年第几月可能比实际日期更重要，比如，月份会对发现数据的季节性趋势很重要。可能需要从日期中提取这类信息，并且把它输入网络作为实际日期的替代或附加。

住址字段或任何文本字段具有相似的复杂性。将地址直接送入网络通常是无用的，虽然能找到一种好办法将整个字段映射到单一值。地址可能包含邮政编码、城市名、州和门牌号，所有这些可能都是有用的特征，但作为一个整体的地址字段就是无用的。

7.7 解释结果

神经网络工具担当解释结果的作用。当估计连续值时，输出常常需要按比例换算回正确的范围。例如，用网络来计算住宅的价值，在训练集中，输出值已经被设定，因此 $103 000映射到－1 而 $250 000映射到 1。如果模型稍后被应用到另外一个住宅并且输出是 0.0，那么我们能够断定其对应值是 $176 500，恰好在最小值和最大值之间。这个逆转换使神经网络估计连续数值变得特别容易。虽然这一步经常不是必需的，尤其是当输出层使用线性转换函数的时候。

对于二元或分类输出变量，方法仍然是把用作训练网络的转换进行逆变换。因此，如果指定“流失”值为 1，“不流失”值为－1，那么靠近 1 的值表示流失，靠近－1 的值表示不流失。当有两个结果时，输出的含义取决于用来训练网络的训练集。因为网络已经学会将误差减到最小，网络训练期间产生的平均值通常接近训练集当中的平均值。解决这个问题的一

个方法是网络找到的第一个模式是平均值。因此，如果最初的训练集有 50% 的流失和 50% 的不流失，那么，网络在训练集实例上产生的平均值要接近 0.0。比 0.0 高的值更像流失，比 0.0 小的像不流失。如果最初训练集有 10% 的流失，那么更合适的分界值应该是 -0.8 而不是 0.0（-0.8 是从 -1 到 1 距离的 10%）。因此，在这种情况下，网络输出看起来确实很像概率。然而，概率在训练集中依赖于输出变量的分布。

还有另一个方法就是对数值赋予置信度。这个置信度会把网络的实际输出看做流失倾向，如表 7-4 所示。

表 7-4　NN 输出的类别和置信度水平

输出值	类别	置信度
-1.0	A	100%
-0.6	A	80%
-0.02	A	51%
+0.02	B	51%
+0.6	B	80%
+1.0	B	100%

对于二元数值，也可以创建产生两个输出的网络，每个对应于一个数值。在这种情况下，每个输出代表“类是正确的”迹象之强弱程度。然后选出的类会是有较高值的那一个，其置信度是基于两个输出强弱程度的某个函数。当两个输出结果不同的时候，这个方法尤其有价值。

提示： 由于神经网络产生连续数值，网络的输出可能难以解释分类结果（在分类中使用）。校正输出的最佳方法是，在验证集上运行网络，完全与训练集分开，并且使用验证集产生的结果来校正网络给出的分类结果。在许多情况下，网络会为每个类分配独立的输出结果，亦即，每个类对应一个倾向。即使是分立的输出，仍然需要用验证集校正输出。

当要考虑两个以上选择项的时候，方法类似。例如，一个长途电信公司尝试瞄准一个新客户集，并提供三种目标服务：

- 对所有的国际呼叫打折扣
- 对非国际的所有长途呼叫打折扣
- 对预先确定的客户集的呼叫打折扣

电信公司是要对三个包中任意一个包的客户提供促销。但由于促销是昂贵的，所以电信公司需要为适当的客户选择正确的服务，以便使活动有好的收益。对所有客户提供全部的三种产品代价是昂贵的，甚至更糟的是，这可能混淆接受者，减少响应率。

电信公司尝试将产品销售给一组客户，他们收到三个产品但只允许对其中的一个做出响应。目的是使用这个信息建立模型预测每个产品对客户的吸引力。训练集使用从销售活动中收集的数据，并且各项编码一般设置如下：无响应→-1.00，国际→-0.33，国内→+0.33，特别号码→+1.00。在使用有关客户信息训练神经网络后，电信公司开始应用模型。

但是，应用模型的过程并不像计划的那样好，许多客户呈簇状围绕在用于训练网络的四个值周围。而且除了无响应者（占大多数）之外，许多情况下，网络返回像 0.0 和 0.5 这样

的中间值。该怎么办呢？

首先，电信公司应该使用验证集来解释输出值。通过用验证集中发生的事情解释网络的结果，它可以找到正确的范围，将网络结果返回到营销部门。图 7-11 所示的就是这样的过程。

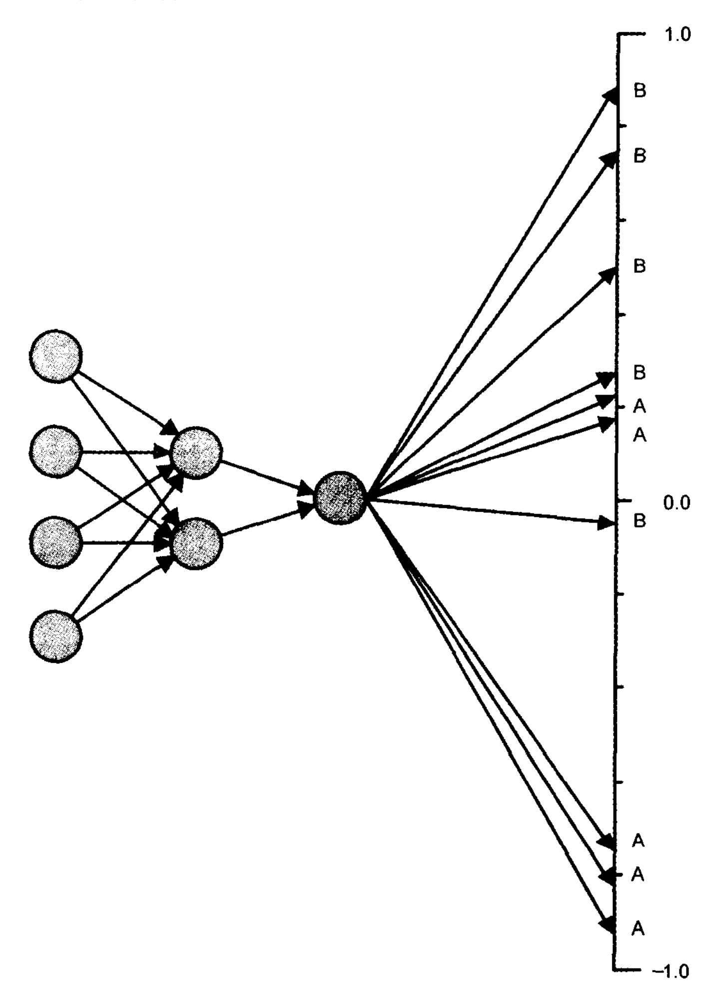


图 7-11 在来自验证集的 10 个实例上运行神经网络有助于确定如何解释结果

对于这种情况，另外一种思路是，真的把网络同时用于预测三种不同的事情，即接受者是否将对每个活动做出响应。这强烈建议我们，网络的较好结构应该有三个输出：对国际计划、长途计划和特定号码计划响应的倾向，然后用测试集来决定无响应者的界限在哪里。另一个可能的选择是，对每个输出分别建模，组合不同模型的结果以选出合适的营销方案。

7.8 时间序列神经网络

在许多商业问题中，数据自然落入时间序列。此类序列的实例包括 IBM 股票的收盘价格、每日瑞士法郎兑换美元的汇率值，或者对未来任何给定日期仍活跃的客户数目的预测。对于金融时间序列，那些能够预测下一个数值或序列是否正在向上或向下发展的人，比其他的投资者有更大的优势。时间序列不止在金融界占尽风头，也可用在其他领域，如预报和过

程控制。但金融时间序列的研究是最深入的，因为预言性能力的小小优势可以转变为很大的收益。

神经网络很容易被时间序列分析所采用，如图 7-12 所示。网络在时间序列数据上进行训练，从数据最古老的点开始，然后移到第二个最古老的点，并且最古老的点在输入层中走到下一个单元集，如此反复进行。像前馈、反向传播网络一样，网络训练尝试在每一步中预测序列的下一个值。

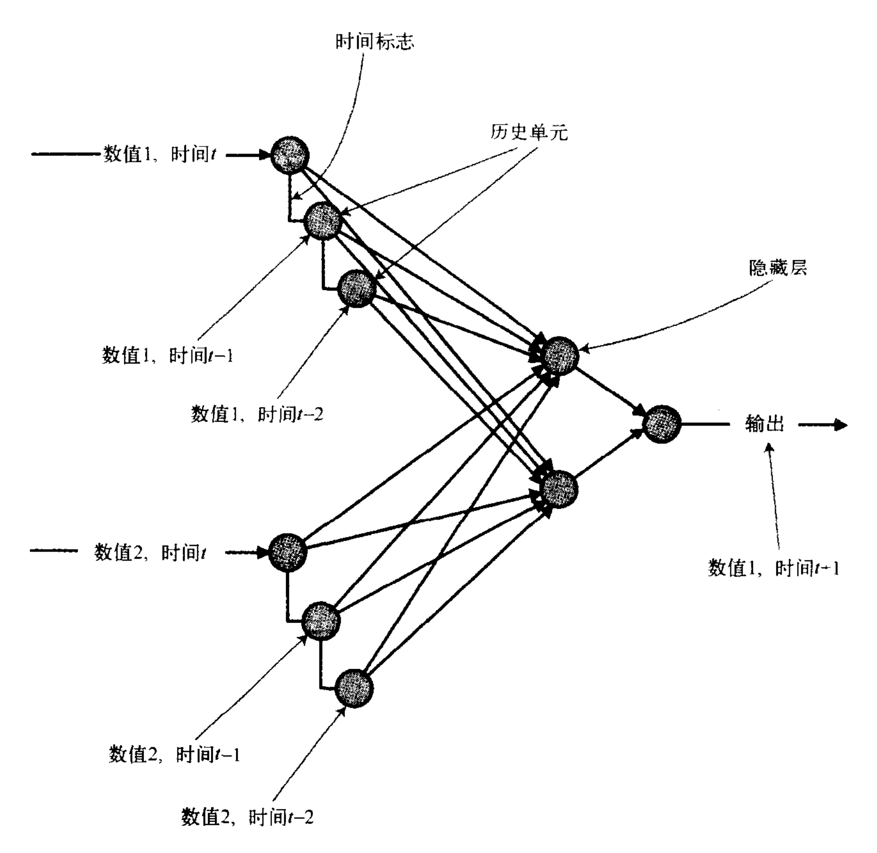


图 7-12 时滞神经网络记住以前少数训练例子，并且用它们作为网络的输入，然后网络像前馈、反向传播网络一样工作

注意，时间序列网络不限于单一时间序列的数据，它能处理多重输入。例如，为预测瑞士法郎兑美元的汇率值，其他的时间序列信息可能被包括在内，如前一日的交易量、日元兑美元的汇率、证券交易所收盘值，以及交易当天是周几，等等。此外，非时间序列数据，如全国范围内调查报告的一段时间内的通货膨胀率可能也是候选的特征。

历史数据的数量控制网络所能识别模式的跨度。例如，在网络上保持 10 个历史单元来预测受欢迎股票的收盘价格，将允许网络识别在未来 2 周内（因为交易价只在工作日提供）发生的模式。但依赖这个网络预测未来 3 个月的数值可能不是个好主意，我们也不推荐这样做。

实际上，通过修改输入，前馈网络能像时滞神经网络一样工作。考虑过去 10 天的历史时间序列，如表 7-5 所示，网络将包括两个特征：周几和收盘价格。

表 7-5 时间序列

日期元素	周 几	收盘价格
1	1	$40.25
2	2	$41.00
3	3	$39.25
4	4	$39.75
5	5	$40.50
6	1	$40.50
7	2	$40.75
8	3	$41.25
9	4	$42.00
10	5	$41.50

创造一个时间延迟为 3 的时间序列需要为历史数据增加新的特征：延迟日的收盘价（周几不需要考虑，因为它的确没有改变）。结果见表 7-6。现在，这个数据可以输入到不需要特殊时间序列支持的前馈和反向传播网络中。

表 7-6 带有时间延迟的时间序列

日期元素	周 几	收 盘 价	上日收盘价	前日收盘价
1	1	$40.25		
2	2	$41.00	$40.25	
3	3	$39.25	$41.00	$40.25
4	4	$39.75	$39.25	$41.00
5	5	$40.50	$39.75	$39.25
6	1	$40.50	$40.50	$39.75
7	2	$40.75	$40.50	$40.50
8	3	$41.25	$40.75	$40.50
9	4	$42.00	$41.25	$40.75
10	5	$41.50	$42.00	$41.25

7.9 如何了解在神经网络内部正在运行的事情

神经网络是不透明的。即使知道遍及网络各处所有结点的所有权重，也不能提供网络为什么产生某个结果的原因。这种理解缺失带有某种哲理性的东西——毕竟，我们不知道人的意识是如何从大脑的神经元产生的。但事实上，不透明损害了理解网络产生的结果的能力。

真希望网络能告诉我们它是如何以规则形式做出决定的。不幸的是，让它如此有力的神经网络结点的非线性特性，同样也使它不可能给出简单的规则。最终，也许提取神经网络规则的研究能带来好的清晰结果。但在此之前，被训练的网络本身就是规则，人们需要用其他方法仔细观察，才能了解网络内部发生的事情。

一项被称为*灵敏度分析*（sensitivity analysis）的技术可用于获知不透明的模型如何工作。灵敏度分析并不提供清晰的规则，但是，它的确暗示了输入对网络结果的相对重要性。灵敏度分析使用测试集决定网络输出相对每个输入的敏感程度。下面是基本的步骤：

1）对每个输入找到平均值。可以把这个平均值当作测试集的中心；

2）当所有的输入在平均值附近时，测定网络输出；

3）当每个输入被修改时，逐个测定网络的输出，确定它的最小值和最大值（通常分别是-1和1）。

如果某些输入使得网络的输出在这三个数值（最小值、平均值和最大值）上改变很少，则网络对这些输入是*不敏感的*（至少当所有的其他输入处于它们的平均值的时候）；而如果输入对网络输出有很大的影响，则网络对其是*敏感的*，可以用对应于每个输入的输出变化量测定网络的灵敏度。对所有输入使用这些方法，就能创造出每个特征重要性的相对度量。当然，这个方法完全是经验性的，并且只独立地观察每个变量。神经网络之所以令人感兴趣，恰恰是因为它们能考虑变量之间的相互作用。

当然在程序方面可以有些变化，可以同时修改两个或三个特征值，以观察某些特征组合是否特别重要。从测试集中点之外的某个位置开始有时是有用的，例如，可能重复分析一些特征的最小值和最大值，以便观察网络对极端情况多么敏感。如果灵敏度分析对这三种情况产生明显不同的结果，那么，在网络中利用特征组合就有更高的优先次序（即需要优先考虑利用这些组合）。

当使用前馈、反向传播网络时，灵敏度分析可以充分利用在学习阶段中计算的误差结果，而不是去独立测试每个特征。将验证集送入网络产生输出，然后将输出与预期的输出相比较计算误差，网络经过单元再把误差传送回来，其目的不是调整任何权重值，而是追踪对应于每个输入灵敏度的轨迹。误差其实就是灵敏度的代言人，利用它可以确定网络中每个输入会在多大程度上影响输出。在整个测试集上，累积这些灵敏度就可以确定哪些输入对输出的影响比较大。但根据我们的经验，这种方式产生的值对了解网络并不是特别地有用。

提示：神经网络不产生容易让人领会的、解释它们如何得出给定结果的规则，但通过使用灵敏度分析，了解网络输入的相对重要性还是可能的。灵敏度测试可以是人工操作的过程，在这个过程中，可以逐一测试每个特征相对于其他特征的灵敏度；通过利用反向传播产生的灵敏度信息，这个过程也可能更加自动化。在许多情况下，理解输入的相对重要性几乎等同于有清晰的规则。

7.10 自组织映像

自组织映像（self-organizing map，SOM）是用于非定向数据挖掘任务（如簇检测）的神经网络变体之一。芬兰研究者Tuevo Kohonen博士发明了自组织映像，所以也被称为Kohonen网络。虽然这些网络原本是用于图像和声音的，但也能识别数据中的簇。它们是以与前馈、反向传播网络相同的基本单元为基础的，但是SOM在以下两个方面与它们完全不同：一是有不同的拓扑结构，此时反向传播的学习方法不再适用；二是有一个完全不同的训练方法。

7.10.1 什么是自组织映像

自组织映像（SOM）是一种能够在数据中识别未知模式的神经网络，图7-13中给出了一个实例。像我们已经看到的网络一样，基本的SOM有一个输入层和一个输出层：输入层的每个单元都连接到一个源，这与建立预言性模型的网络一样；同时，像其他网络一样，SOM中的每个单元都有一个独立的权重，与每个进入的连接相关联（实际上，这是所有神

经网络的特性)，但 SOM 与前馈、反向传播网络之间的相似性也仅此而已。

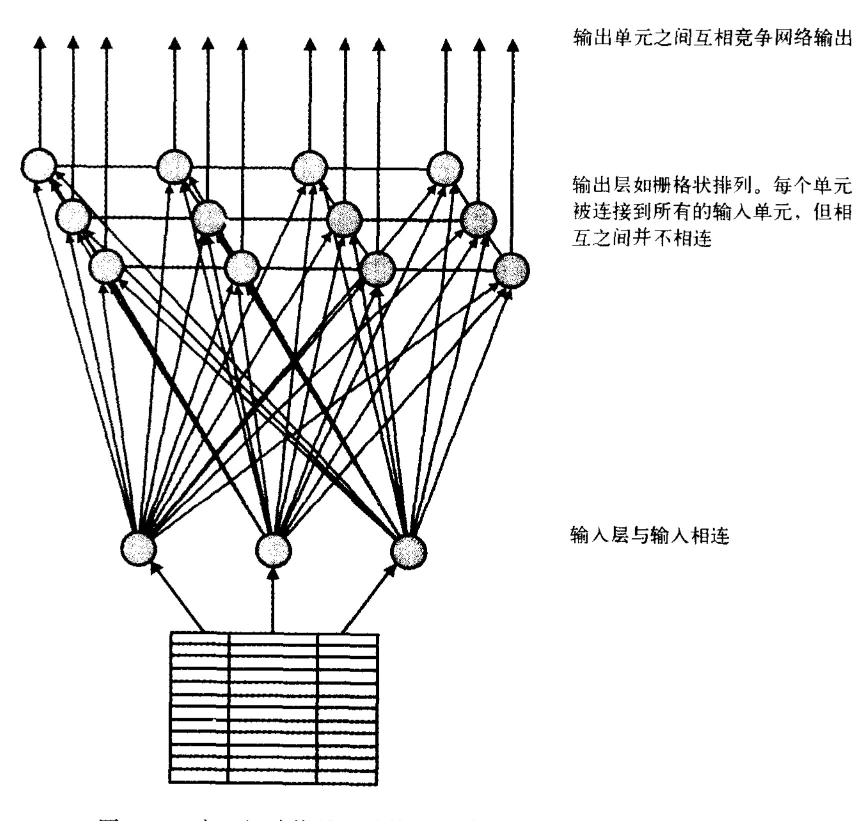


图 7-13　自组织映像是一种特别的神经网络类型，能用来发现簇

输出层是由许多单元构成的，而不是少数几个。输出层中的每个单元都连接到输入层的所有单元。输出层呈栅格状排布，像一个棋盘格。即使在这个层中的单元彼此并不相连，栅格状的结构在 SOM 训练中实际上扮演着重要角色，我们稍后会看到。

SOM 是如何识别模式的呢？设想在狂欢节上有一个亭子，你在那里朝布满孔洞的墙上投掷球。如果球落入其中的一个洞，你可以选择奖品。训练 SOM 就像一个人蒙着眼处于亭子中，并且最初墙上没有洞，这种情况与下面描述的情况非常相像：开始在大量的数据中寻找模式，但不知道该从哪里下手。当你每次投掷球后，墙上便留下小的凹痕，最后，当足够多的球投在相同区域附近时，凹陷穿破墙形成一个洞。现在，当另外一个球落在那个位置时，便穿洞而过，你将获得一个奖品——在狂欢节上，它是一个廉价的毛绒玩具，而在 SOM 中，是一个可以确认的簇。

图 7-14 显示了一个简单的 SOM 是如何运转的。当许多的训练集用于训练网络的时候，数值经过网络向前流到输出层的单元。输出层的单元之间彼此竞争，有最高值的那一个“胜出”。获得的奖赏是调整通向获胜单元路径的权重，强化对输入模式的响应。这就像在网络中产生一个小凹痕。

对网络的训练还有另外一个方面。不仅获胜单元的权重被调整，而且紧邻它的单元的权重也被调整，以强化它们对输入的响应。这种调整由邻近度参数来控制，这个参数可以控制

邻居的数目和调整量。最初，邻居的数目相当大，并且调整量也很大。当训练继续进行，邻居的数目和调整量开始减少。邻近度参数实际有以下几个作用：一是输出层表现更像相互联系的纺织物，虽然单元之间彼此并不直接相连；与那些不相似的簇相比，相互之间相似的簇应该靠得更近；更重要的是，邻近度参数可以把一组单元表示为单个簇，如果没有邻近度参数，网络一般会找到与数据中输出层单元数目一样多的簇——这就在簇检测中引入了偏离。

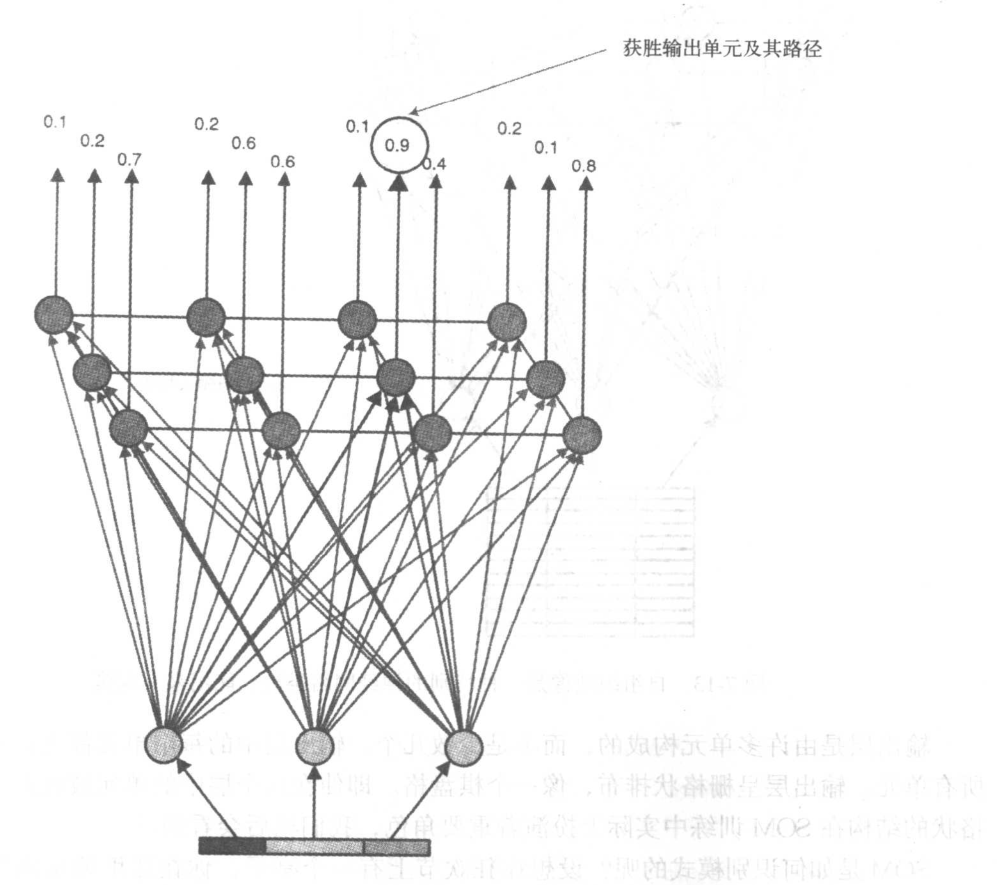


图 7-14　一个 SOM 能找到出色地识别特定输入的输出单元

比较典型的情况是，SOM 可以识别出少于输出单元数目的簇。当使用网络分配新的记录到簇中的时候，其效率是很低的，这是由于新的输入经过网络馈入到输出层中从未使用过的单元。为了确定实际上使用了哪个单元，我们将 SOM 应用到验证集。验证集成员被送入网络，追踪每种情况下胜出单元的轨迹，那些没有采用或很少采用的单元被丢弃。去掉这些单元可以增加网络运行时的表现，因为这可以减少对于新实例的运算次数。

一旦最终网络就位，即输出层中只含有那些能够识别特定簇的单元，网络就可以应用于新例子。一个未知的例子输入到网络中，被分配到输出单元中有最大权重的簇。网络已经识别出了簇，但是我们不知道与它们相关的任何情况。稍后我们将回到识别簇的问题上。

最初的 SOM 使用二维栅格作为输出层，它是早期为识别由二维像素值阵列组成的图像特征而构造出来的。输出层实际可以是任意结构——可以是在三维空间中定义的邻居，如六角形的网络或其他形式的布局。

7.10.2 实例：发现簇

一家大银行对增加正在推销的住宅抵押贷款项目的数量非常感兴趣，这提供了一个聚类的实例。这家银行决定，要了解目前的住宅抵押贷款客户的情况，以确定最佳策略，增加市场占有率。为启动这个过程，他们收集了购买住宅抵押贷款的 5 000 个客户和没有购买产品的 5 000 个客户的人口统计学数据。即使持有住宅抵押贷款的客户比例少于 50%，在训练集设置相等的权重仍是一个不错的主意。

所收集的数据有如下字段：

- 住宅评估值
- 有效的信用额度
- 允许的信用额度
- 年龄
- 婚姻状况
- 孩子数目
- 家庭收入

这个数据构成了一个好的聚类训练集。输入值被映射到 -1 和 +1 之间，然后用于训练 SOM。网络在数据中识别出了五个簇，但不提供有关簇的任何信息。这些簇到底意味着什么呢?

用于比较神经网络技术中运行特别好的簇的常用技术是*平均成员技术*，找到每个簇的最平均的成员——簇的中心。这与灵敏度分析的方法相类似。为达到这个目的，首先要找到每个簇中每个特征的平均值。由于所有的特征都是数字，对于神经网络来说，这不成问题。

例如，假定簇的一半成员是男性，另一半是女性，并且男性映射到 -1.0，女性映射到 +1.0，则该簇的平均成员的这个特征会有 0.0 值。在另一个簇中，可能有 9 个女性，1 个男性，对于这个簇，平均成员会有 0.8 的值。因为所有的输入都必须被映射到一个数值范围之内，这种取平均值的方法在神经网络上运行效果良好。

提示： *自组织映像属于神经网络的一类，可以用于识别簇，但是不能指出到底什么使得簇成员之间彼此相似。一个可用于比较簇的有力方法是，在每个簇中确定中心成员或平均成员。使用测试集，计算数据中每个特征的平均值，然后，这些平均值显示在同一个图表中，可以确定出簇的独有特征。*

接下来，这些平均值可以用图 7-15 所示的平行坐标作图，它给出了在金融业实例中识别出的五个簇中心。在这种情况下，银行注意到其中一个簇特别值得关注，该簇由四十岁左右、有孩子的已婚客户组成。进一步的调查显示，这些客户的孩子年龄都在十八、九岁，与其他簇的成员相比，这个簇的成员有比较多的住宅抵押贷款。

故事继续发展，银行的销售部门得出结论，这些人用住宅抵押贷款支付孩子们上大学的学费。该部门准备专门为这个市场设计营销计划，通过销售住宅抵押贷款的方法支付大学的教育费用。而这项活动的结果令人失望，营销计划是不成功的。

由于营销计划的失败，好像簇没有实现它们的承诺。事实上，问题出在其他地方，银行最初只是使用一般客户的信息，没有通盘考虑所服务的众多不同渠道的客户信息。银行回到了识别客户的问题上，但是这次包括了较多的信息——来自存款系统、信用卡系统等的信息。

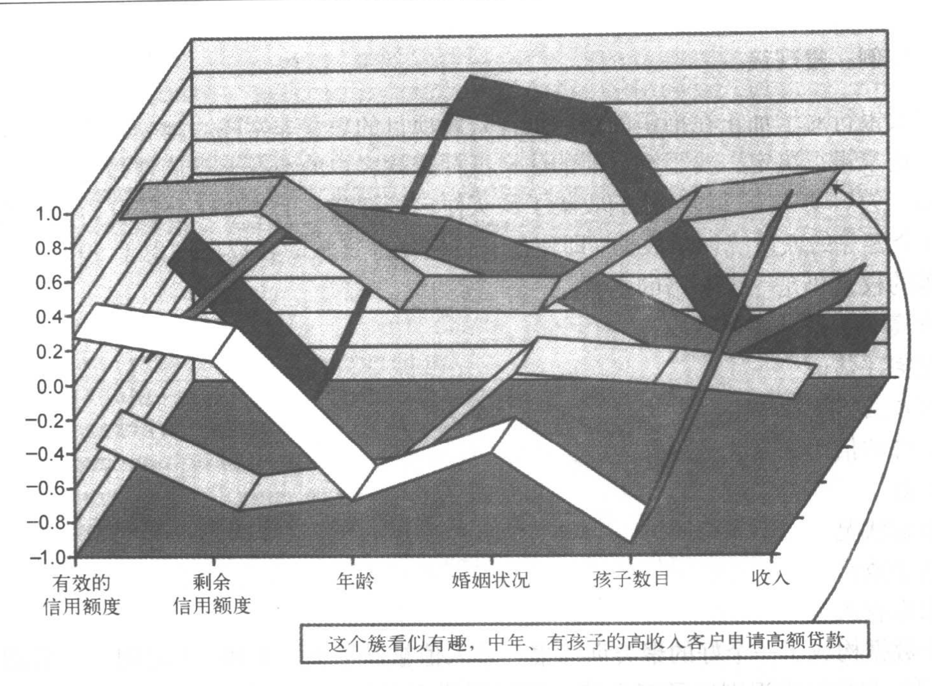


图 7-15 在同一个图上比较五个簇中心。这项简单的可视化技术（被称为平行坐标）帮助识别有趣的簇

基本方法是一致的，因此，我们将不再详细讨论有关的分析。利用后来附加的数据，银行发现，有适龄上大学孩子的客户簇实际上是存在的，但是这个事实被忽略了。当包括附加数据时，银行获得的信息是，该簇的客户除了有个人账户外，也往往有企业账户。这带来了值得思考的新问题：当孩子离家去上大学的时候，这些父母就有机会利用家中的债券开始新的商务活动。

利用这种洞察力，银行设计出了专门针对这类父母的营销计划，主要是关于在“空巢”时开展新商务的计划。这个计划获得成功，而且银行看到了住宅抵押贷款的客户组表现良好。从这个案例研究获得的经验是，尽管 SOM 是发现簇的有力工具，事实上神经网络的能力最多只能与输入其中的数据达到同样好的水平。

7.11 小结

神经网络是用途广泛的数据挖掘工具。在很多行业和大量的应用中，神经网络一再证明了自己的重要性。这些来自于复杂领域的结果，例如分析时间序列和发现欺诈行为，不是其他技术容易做到的。目前最大的神经网络产品或许是 AT&T 为在支票上读数字而开发的系统，这个神经网络将数十万计的单元编入七个层中。

神经网络是基于大脑如何工作的生物模型而创建。尽管使用的是早期的数字计算机，但基本理念已被证明是非常有用的。在生物学上，神经元在输入达到一定的阈值（threshold）后被激发，这种模型也能被移植到计算机上。实际上该领域从 20 世纪 80 年代才蓬勃发展起来，当时统计学家开始使用它们，并且更好地了解它们。

神经网络由相互关联在一起的人工神经元组成。每个神经元都模仿它的对应生物体，利

用不同的输入组合产生输出。由于数字神经元处理的是数字，激活函数就代表了神经元的特征。在大多数情况下，该函数是取其输入的加权和，然后再应用S形函数处理它。结果是，一个结点有时以线性方式出现，有时以非线性方式出现——这是在标准统计技术上的改进。

最常见的神经网络是用于预言性建模的前馈网络，但最初的突破点——反向传播训练方法——已经被其他方法替换，例如“共轭梯度”。这些网络既能应用于分类型输入，也能应用于连续型输入，但只有当输入字段被映射到-1到+1的范围时，神经网络能获知的才最多，这是有助于训练网络的一个指导性原则。当少量的数据落在范围之外，而且范围更窄（比如0到1）时，神经网络仍然可以运行。

神经网络确实有几个缺点。首先，当只有少数几个输入变量的时候，它们运转最好，但技术本身无助于选择使用哪些变量，变量选择可以利用其他技术（如决策树）帮助解决；同时，在训练网络时，不能保证产生的权重组是最佳的，为了增加结果的置信度，可以建立几个网络，选定其中的最佳者。

也许最大的问题是，神经网络不能解释它正在做什么。决策树很常用，因为它们能提供一系列规则，但从神经网络中不可能获得准确的规则组。神经网络只能由它的权重和非常复杂的数学公式做出解释，不幸的是，这个问题超出了人类的理解能力。

神经网络的变体，如自组织映像，可以把该技术扩展到非定向聚类。总体来说，神经网络是强有力的，能够产生好的模型，不足之处是我们不知道它们是如何工作的。

第8章　最近邻方法：基于存储的推理和协同过滤

你听到某人说话，立刻就会猜测她来自澳洲，为什么呢？因为她的口音使你回想起你曾经遇见的其他澳洲人。又比如，受一个吃饭比较讲究的朋友推荐，你准备到一家有可能喜欢的新餐馆用餐。上述这两种情况都是以经验为基础进行判断的例子。当面对新形势时，人们很自然地被过去曾经历的类似情形的记忆所引导，这就是本章要讲的数据挖掘技术的基础。

最近邻技术正是基于这种"相似性"概念。基于存储的推理（memory-based reasoning，MBR）的结果以过去类似的情形为基础——非常像基于过去所知道的澳洲口音来判断一个新朋友是澳洲人这个例子。协同过滤会增加更多的信息，因为它不仅使用邻居之中的相似性，还同时考虑他们的不同喜好。餐馆推荐就是一个协同过滤（collaborative filtering）的实例。

所有这些技术的中心是"相似性"这个概念。到底是什么造成了过去发生的事件与一个新的事件相似？在从过去寻找类似的记录的同时，还要想办法把邻居的信息结合起来。这就是最近邻方法的两个主要概念。

本章将首先对MBR进行简单介绍，解释它是如何工作的。对最近邻技术而言，距离和相似性的衡量是很重要的，所以后面有一节专门讲述距离度量，包括不同数据类型距离的意义（比如纯文本中距离就没有明显的几何学解释）。

MBR的思维方法可以透过一个案例表现，这个案例讲述了MBR如何把关键词穿插到新闻报导中去。本章最后将讲到协同过滤，这是一个做出推荐时常用的方法，尤其在网上更常用。协同过滤也是以最近邻方法为基础的，但是有微小的改变——它不是通过把餐馆或电影分成不同邻居组，而是按照推荐餐馆或电影的人来分组。

8.1　基于存储的推理

人们从经验推理的能力依赖于从过去找到合适样本的能力。医生诊断疾病，理赔分析员标识出欺诈保险索赔，采蘑菇的人发现羊肚菌，这些活动都遵循一个相似的过程，即每个人首先从经验中找出类似的案例，然后把他们从这些案例中得到的知识应用于需要解决的问题。这就是基于存储的推理之精髓所在：从一个已知记录数据库中搜寻与一条新的记录相类似的预分类记录，然后把这些邻居记录应用于分类和估计。

MBR的应用横跨许多领域：

欺诈探测：新的欺诈案例可能与已知的案例类似，MBR识别出它们并加以标记，以便做进一步的调查。

客户响应预测：下一个有可能响应优惠服务的客户，或许与以前已经响应的客户类似，MBR能容易地识别出下一个可能的客户。

医学治疗：对一位现有病人最有效的治疗或许是对其他类似病人达到最佳效果的治疗方法，MBR 能发现达到最佳效果的治疗方法。

把响应分类：像美国人口普查表上的那些职业和行业，或者客户的抱怨这样的文本响应，需要被归入一系列固定的分类代码，MBR 能处理这些自由文本并分配代码。

MBR 的强有力的特点之一是它“原样”使用数据的能力。与其他的数据挖掘技术不同，它不关注记录的格式，只关注两种运算：一是*距离函数*，用于计算在任意两个记录之间的距离；二是*组合函数*，用于结合几个邻居的结果形成一个答案。这些函数是针对许许多多记录而定义的，包括复杂的或不常用的数据格式记录，如：地理位置、图像和自由文本等通常难以用其他的分析技术处理的数据。本章稍后将讲到的一个案例展示了 MBR 在新闻报导分类方面的成功应用——一个充分利用新闻报导的自由文本分配代码的实例。

MBR 的另一个强有力的特点是它的适应能力。只需要将新的数据纳入历史数据库，MBR 就能利用它从旧的类或定义中得到新的类和定义。MBR 不需要一段很长的时间来训练或更改数据为某个正确的格式，也能照样给出好的结果。

这些优势是有代价的。MBR 往往是数据贪婪者，为了找到邻居，需要使用大量的历史数据，为分类新记录，需要处理所有的历史记录以找到最相似的邻居——与已经训练过的神经网络或已建立的决策树等方法相比，是一个更耗时的过程。另一个必须面对的问题就是发现好的距离函数和组合函数，这通常需要一些尝试，有时可能出现错误，当然还需要有某种直觉等。

实例：使用 MBR 估计纽约州 Tuxedo 镇的房租

这个实例的目的在于说明 MBR 是如何工作的，通过结合几个相似城镇（目标城镇的最近邻）的租金数据，估计目标城镇中一套公寓的租赁费用。

MBR 过程首先要找出最近邻，然后把它们的信息结合起来。图 8-1 说明了这些步骤中的第一步。任务目标是通过察看纽约州 Orange 县 Tuxedo 镇的最近邻情况，来预测它的房租。所谓的最近邻不是指位于纽约州东部的 Hudson 和 Delaware 河流那些*地理上的*邻居，而是基于描述变量的邻居——在这个案例中，指的是人口和中值住宅价格。散点图给出的是按这两个变量画出的纽约城镇情况，从图 8-1 可以看出，以这样的方式来看，Brooklyn 和 Queens 是最近的邻居，两者离曼哈顿都很远。虽然曼哈顿的人口密度几乎与 Brooklyn 和 Queens 一样，但由于住宅价值，它被单独列为一个类。

> **提示**：*邻居的概念可以从各个角度来理解。角度的选择决定哪些记录是彼此接近的。对于某些目的而言，地理位置的接近可能很重要；对于其他目的来说，住宅价值或平均住宅面积大小或人口密度（density）可能是更重要的。角度的选择和距离度量的选择对任何最近邻方法都是非常重要的。*

MBR 的第一个阶段是在图 8-1 中所示的散点图上寻找最近的邻居，然后再找到下一个最近的邻居，如此反复直到找到所需要的数目为止。在这个例子中，邻居的数目是 2，最近的是 Shelter 岛（是一个真正的岛），其出口是 Long Island 的 North Fork 和 North Salem（Northern Westchester 的一个城镇，靠近 Connecticut 州边界）。这些城镇处于人口排序列表的中部，如果按照住宅价格排序列表，它们位于列表顶端附近。虽然空间距离相隔很多英里，而且位于两个方向，但 Shelter 岛和 North Salem 与 Tuxedo 镇的情况是很类似的。

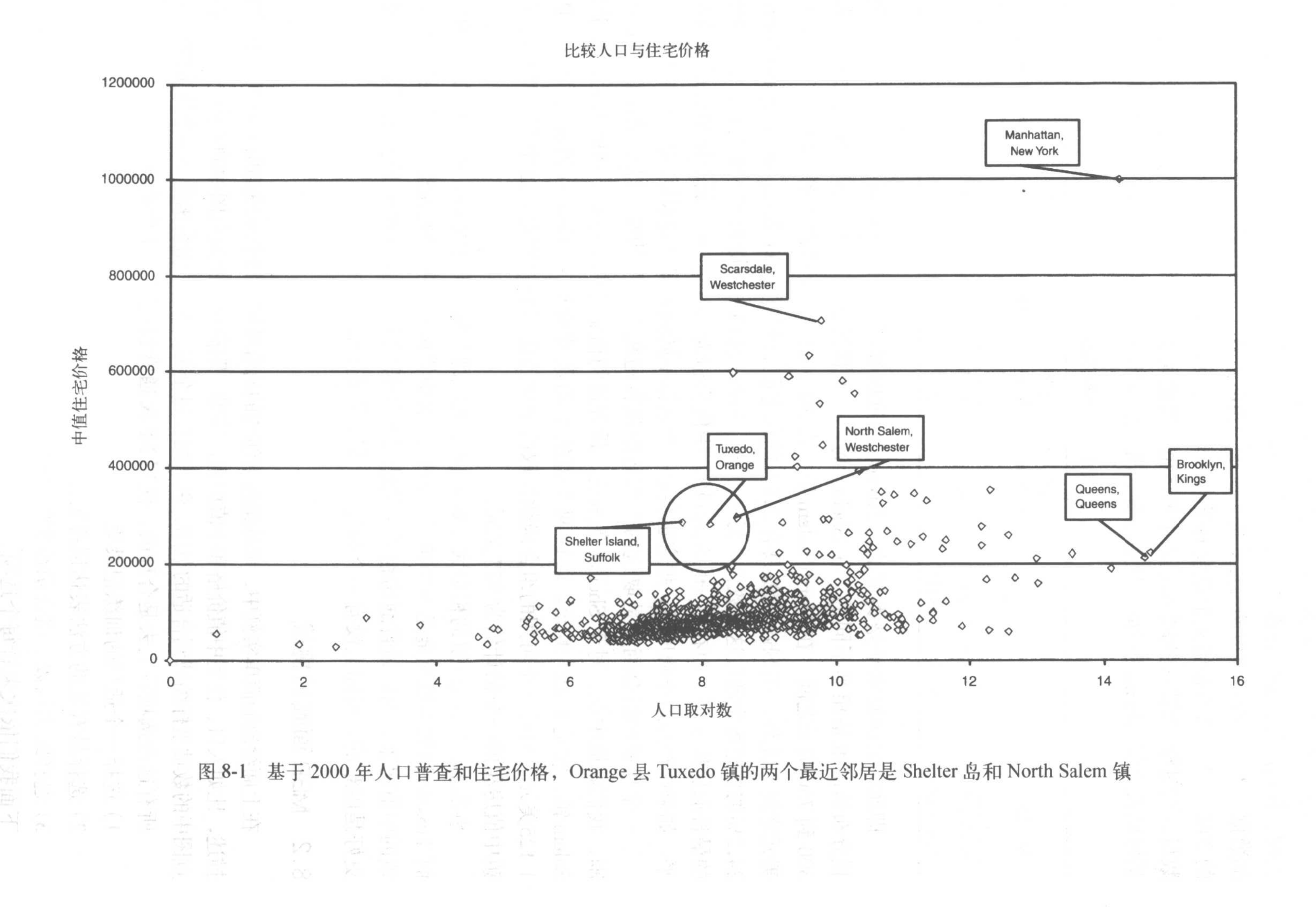


图 8-1 基于 2000 年人口普查和住宅价格，Orange 县 Tuxedo 镇的两个最近邻居是 Shelter 岛和 North Salem 镇

邻居一经确定，下一步是把来自邻居的信息组合起来，推出关于目标的某些事。对于这个实例，我们的目标是估计在 Tuxedo 镇租用住宅的费用。可能有不止一种合理的方法组合邻居的数据。人口普查可以以两种形式提供关于房租的资讯。表 8-1 显示了这两个邻居城镇中选出的 2000 个人口普查报告关于租金的报告结果。对于每个城镇，给出了几个不同价位段的租户数目，同时给出的还有每个城镇的中值房租。关键问题是，找出一个方法能够使用这个数据来最好地表达邻居房租的特征，然后组合邻居信息给出一个表示 Tuxedo 镇的房租特征的估计。

表 8-1 一些邻居

城 镇	人口	中值房租	房租 < \$500 (%)	房租 \$750 (%)	房租 \$1500 (%)	房租 \$1000 (%)	房租 > \$1500 (%)	无房租 (%)
Shelter 岛	2228	\$804	3.1	34.6	31.4	10.7	3.1	17
North Salem	5173	\$1150	3	10.2	21.6	30.9	24.2	10.2

即使中值房租水平是类似的，Tuxedo 最近的邻居——North Salem 镇和 Shelter 岛的房租分布情况也是相当不同的。在 Shelter 岛，一个普通的住宅（占 34.6%的比例），租金在 500 到 750 美元之间。在 North Salem 镇，占 30.9%的最大多数住宅，租金在 1000 到 1500 美元之间。此外，在 Shelter 岛只有 3.1%的住宅租金超过 1500 美元，而在 North Salem 镇有 24.2%的住宅租金超过这个数目。另一方面，Shelter 岛的中值房租为 804 美元，高于 750 美元的最普通房租水平，而在 North Salem 镇，中值房租为 1150 美元，低于该镇的最普通房租水平。如果能够知道平均房租，那它也会是一个表征不同城镇房租的很好的候选参数。

一个可能的组合函数是取这两个邻居的最普通租金的平均数。既然给出的只是一个范围，我们就取中点。对于 Shelter 岛来说，最普通的租金范围中点是 1 000 美元，对于 North Salem 镇，它是 1 250 美元。取这两个值的平均数就给出了 Tuxedo 镇房租的一个估计值 1 125美元。另外一个可用的组合函数是取两个中值房租的中间点，这个方法给出的 Tuxedo 镇中值房租的一个估计值为 977 美元。

事实上，Tuxedo 镇的租金大多数是在 1 000 到 1 500 美元之间，中间点在 1 250 美元，而 Tuxedo 镇的中值租金是 907 美元。所以，取中值房租的平均值就稍微高估了一点 Tuxedo 镇的中值房租，而取最普通房租的平均值又稍微低估了 Tuxedo 镇的最普通租金。说哪一个更好是困难的，实际上没有一个明显的“最佳”组合函数。

8.2 MBR 面临的挑战

在上面给出的简单实例中，训练集包含纽约的所有城镇，每个城镇都用一系列数值型字段来描述，比如人口、住宅中值价格和中值房租，等等。距离可以由散点图中的不同方位来确定，散点图中的数轴坐标缩减到合适的范围，邻居的数目全都定为 2。组合函数是一个简单的平均。

所有这些选项看上去是合理的。使用 MBR 通常包括以下几个选项：

1）选择一个适当的训练记录集。

2）选择最有效的方法表达训练记录。

3）选择距离函数、组合函数和邻居的数目。

下面我们依次来看每个选项。

8.2.1 选择一组平衡的历史记录

训练集是一组历史记录，需要涵盖人口状况，以便一个未知记录的最近邻居可以用于预言性目的。一个随机样本不可能提供对所有数值的充分覆盖：一些类比其他类更常出现，而且更常出现的类往往会占随机样本的大部分。

例如，欺诈交易比非欺诈交易要少许多，心脏病比肝癌发生率更高，关于计算机工业的新闻报导比塑料工业更多，等等。为了达成平衡，如果可能的话，训练集包含的记录应该满足这样的要求：每个类有大约相等数目的记录。

提示： 当为MBR选择训练集的时候，要保证每个类大约有相同数目的记录支持它。依通常的经验判断，要保证适度支持，每个类有数十个记录是一个最低要求，而数百或数以千计的样本都是很平常的。

8.2.2 表示训练数据

MBR在预测方面的表现依赖于训练集的表示。图8-2所示的散点图方法对于两、三个变量和小的记录数目是有效的，但是不能很好地扩展。找到最近邻居的最简单方法是，找出未知事件与训练集中每一个记录的距离，并选出距离最小的训练记录。当记录数目增加的时候，为一条新记录寻找邻居需要的时间会增加地很快。

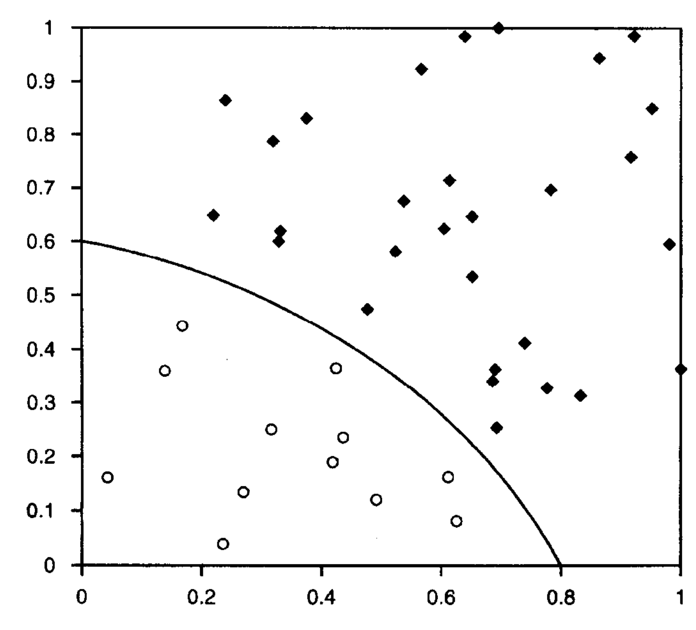


图8-2 大概MBR的最简洁的训练集是恰好整齐地分为两个不相交的集合

在记录被储存在一个关系型数据库时尤其如此。对于本案例，需要用到的查询类似如下：

```
SELECT distance(), rec.category
FROM historical_records rec
ORDER BY 1 ASCENDING;
```

记号distance()中的值可填写为一个特定距离函数。在本案例中，要找到少数几个最近的

邻居，需要给所有的历史记录排序。这需要全表扫描，再加上排序——两个相当费劲的运算。我们可以通过遍历表格，保存最近邻居的另一个表，视情况插入和删除记录，这样就可以不用排序了。不幸的是，如果不使用一种程序语言，这个方法在 SQL 中是不容易完成的。

关系数据库目前的表现相当好。为 MBR 数据评分面临的挑战是，每一个待评分的事件都需要与数据库中的每一个事件进行对比，即使有数以百万计的历史记录，给单一的新记录评分并不需要花费很多时间；然而，同时给许多新记录评分可能效果会比较差。

提示 MBR 效率的另外一个方法是减少训练集中的记录数目。图 8-2 显示了分类数据的一个散点图。在这个图中的两个区域之间，有一条明确的分界线。在线上面的所有点是菱形点，而所有线下面的点是圆形点。虽然这个图中有 40 个点，但是大部分是多余的。也就是说，它们并不是满足这个分类目的所必需的点。

图 8-3 表明，只需要 8 个点就可以得到相同的结果。由于训练集的大小对 MBR 的表现有如此大的影响，减少它的大小会极大地增强 MBR 的性能。

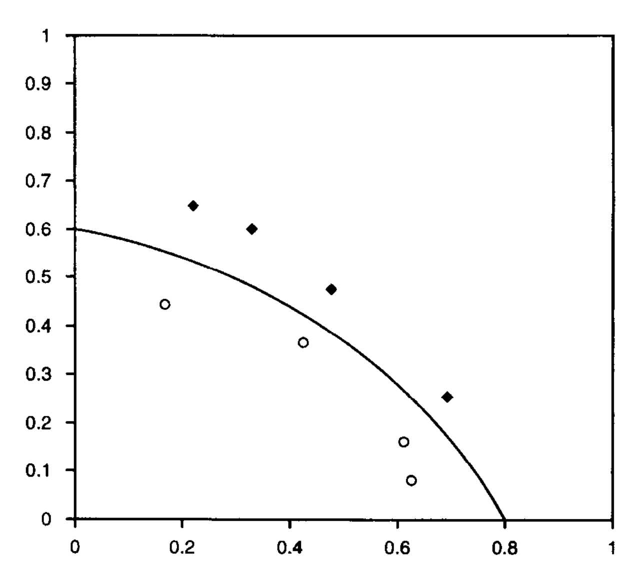


图 8-3 使用更少的点集给出与图 8-2 中 MBR 相同的结果

如何才能找到这个缩小的记录集合？最实用的方法是找寻包含属于不同种类的记录的簇，然后把这些簇的中心部分作为一个缩减的集合使用。当不同的类是分散的时候，效果很好；然而，当有一些重叠，而且类定义不太明确的时候，使用簇减少训练集的大小会导致 MBR 给出很差的结果。发现最佳“支持记录”的集合已经成为近期研究的一个热点。找到一个最佳集合的时候，历史记录有时能被缩减到可以填入一个电子表格那样大小，这样，在计算能力比较差的机器上把 MBR 应用到新的记录时会更有效。

8.2.3 确定距离函数、组合函数和邻居的数目

距离函数、组合函数和邻居的数目是使用 MBR 时要用到的几个关键部件。根据判别标准的不同，同一组历史记录对于预言性目的可能很有用，也可能一点用处都没有。幸运的是，简单的距离函数和组合函数通常可以相当好地满足工作需要。在详细讨论这些问题之

前，先来看一个详细的案例介绍。

8.3 案例研究：分类新闻报导

这一案例研究使用 MBR 为不同的新闻报导分配类别代码（classification code)，是以本书的一位作者指导的工作为基础的。这个案例结果显示：MBR 也能与人一样在解决涉及数百个类和难以使用的数据类型（如自由文本）的问题上工作良好[⊖]。

8.3.1 什么是代码

类别代码是用来描述新闻报导内容的关键词。这些代码由新闻检索服务（news retrieval service）添加到报导中，可以帮助使用者寻找感兴趣的报导。它们帮助把关于某些特别事件的报导自动发送给特别客户，而且帮助实现个性化的描绘。例如，汽车工业分析师（或其他任何对该主题感兴趣的人）可以通过找寻含有“汽车工业”代码的文件来简化搜索。因为经验丰富的专家（也称之为编辑）建立了这种代码，人们就可以检索到正确的报导。传统上这些代码是由编辑或专家系统已经分配好的。本案例探索了 MBR 在这一领域的应用。

用于这一研究的代码分为六个种类：

- 国家机关
- 工业界
- 市场领域
- 产品
- 区域
- 科目

这些数据包含了 361 个独立代码，分布在表 8-2 所示的训练集中。

表 8-2 用于分类新闻报导的六种类型的代码

种　类	代码数目	文件数目	出现次数
政府（G/）	28	3 926	4 200
工业界（I/）	112	38 308	57 430
市场领域（M/）	9	38 562	42 058
产品（P/）	21	2 242	2 523
区域（R/）	121	47 083	116 358
科目（N/）	70	41 902	52 751

对不同报导所赋予代码的数目和类型是各不相同的。几乎所有报导都有区域和科目代码——平均来说，几乎每个报导包含三个区域代码。另一个极端情况是，只有极少数报导包含了国家机关和产品代码，而且这类报导很少有一个以上的这类代码。

8.3.2 应用 MBR

这一部分内容将解释 MBR 如何为一个新闻服务社轻松地分配新闻报导代码。包括的几

⊖ 本案例是本书的一位作者指导的一个调查概要。完整的详细内容见文章“利用基于存储的推理分类新闻报导”，作者 David Waltz、Brij Masand 和 Gordon Linoff，SIGIR 会议论文集，1992，ACM 出版公司出版。

个重要步骤是：

1）选择训练集。

2）决定距离函数。

3）选择最近邻居的数目。

4）决定组合函数。

下面的几个小节将依次讨论上述步骤。

1. 选择训练集

由新闻检索服务机构为这一目的提供的训练集包含 49 652 个新闻报导，这些报导来自大约三个月的新闻和几乎 100 个不同的渠道。平均每个报导包含 2 700 个单词，有 8 个指定的代码。训练集并不是特别建立的，因此训练集中代码出现的频率差别很大，基本上可以重现新闻报导中代码的总体频率。虽然这个训练集产生了很好结果，但是建立一个更好的训练集，使其包含更多较不常出现的代码的样本，或许 MBR 的表现会更好。

2. 选择距离函数

下一步是选择距离函数。在本案例中，已经存在一个距离函数，以测量两个文档中所包含单词的相似性的相关性反馈（relevance feedback）概念为基础。相关性反馈概念在下面的"使用相关性反馈创建距离函数"部分有更详细的描述，它最初是用于返回一个给定文档的相似文档，是作为细化搜索的一种手段。最类似的文档即是 MBR 所使用的邻居。

3. 选择组合函数

下一个重要问题是组合函数。给新闻报导分配类别代码与绝大多数分类问题有一点不同。绝大多数分类问题是寻找单一的最佳解决方案。然而，即使在种类相同的情况下，新闻报导也可以有多种代码。MBR 适合解决这个问题的能力更突显了它的灵活性。

使用相关性反馈创建距离函数

相关性反馈是允许使用者基于文本数据库改进搜索的强有力手段，这种改进是通过要求数据库返回类似文档实现的。网络中心和权威（hub and authority）是另外一个在超链接的网页上改良搜寻结果的方法，将在第 10 章中做详细介绍。在相关性反馈过程中，文本数据库中所有的文件都被评分，然后返回那些最相似的文档，同时给出相似性的程度大小，即相关性反馈得分，它可以作为 MBR 距离测量的基础。

在本案例中，相关性反馈得分的计算如下：

1）像"它"、"和"及"的"这样的常用但不具有明确含义的词都被从训练集的所有报导文本中去掉。这一类中总共有 368 个字被识别而去掉。

2）另外一些是最常用的词语，对应于数据库中 20% 的词组，也被从文本中去掉。因为这些词组太平常了，极少能提供区别两个文档的有用信息。

3）剩余的词组被收集进一个可搜寻术语的字典。每个词语被分配一个权重，它反比于在数据库中的出现频率。这个特定的权重是由该术语在训练集中的出现频率取以 2 为底的负对数得到的。

4）以大写字母打头的词组对（如"United States"和"New Mexico"）被自动地识别，包含在可搜寻术语的字典中。

5）为了计算两个报导的相关性反馈得分，将两个报导中可搜寻术语的权重相加。当可搜寻术语在两个报导中表现得极其相近的时候，本案例的算法就给出一个附加分。

相关性反馈得分是改编一个已知函数用作距离函数的实例。然而，得分本身并不完全符合距离函数的定义。特别是，得分为 0 表明两个报导没有共同的词组，而不是暗示报导是恒等的。下列变换把相关性反馈得分转变为适于测量新闻报导之间“距离”的一个函数：

$$d_{classification}(A, B) = 1 - score(A, B) / score(A, A)$$

这是一个用于发现最近邻居的函数。由于 d（A，B）与 d（B，A）不同，实际上这也不是一个真实的距离函数，但是使用它来工作已经可以得到足够好的结果。

组合函数采用了加权和技术。由于最大的距离是 1，权重只是 1 减去距离，因此距离小的邻居权重会大，而距离大的邻居权重就小。举例来说，假设某个报导的邻居有表 8-3 中所示的区域代码和权重。

表 8-3　未分类报导中的邻居分类

邻　居	距　离	权　重	代　码
1	0.076	0.924	R/FE，R/CA，R/CO
2	0.346	0.654	R/FE，R/JA，R/CA
3	0.369	0.631	R/FE，R/JA，R/MI
4	0.393	0.607	R/FE，R/JA，R/CA

一个代码的总得分是包含该代码的邻居的权重总和，得分低于某一个阈值的代码将被除去。例如，代码 R/FE（是远东区域的代码）的得分是邻居 1，2，3 和 4 的权重之和，因为它们全部都包含 R/FE，这样产生一个 2.816 的得分。表 8-4 显示了被至少四个邻居之一包含的 5 个区域代码的得分结果。对于这些例子，1.0 阈值只留下了三个代码：R/CA、R/FE 和 R/JA。特别选取的阈值是基于不同数值进行的实验，对理解 MBR 并不重要。

表 8-4　未分类报导中的代码得分

代　码	1	2	3	4	得　分
R/CA	0.924	0	0	0.607	1.531
R/CO	0.924	0	0	0	0.924
R/FE	0.924	0.654	0.631	0.607	2.816
R/JA	0	0.654	0.631	0.607	1.892
R/MI	0	0.654	0	0	0.624

4. 选择邻居的数目

本案例中，最近邻居的数目只在 1 和 11 之间改变。使用较多的邻居可以得出最佳结果。然而，本案例不同于很多 MBR 应用实例的地方是，它给每个报导分配多个类。比较典型的问题通常是只分配单个类或者代码，而且较少的邻居通常就可以产生足够好的结果。

8.3.3 结果

为了测量 MBR 对分配代码的有效性，新闻服务社派一个编辑小组复审对所有 200 个报导的代码分配，其中包括通过编辑和 MBR 方法分配的。只有那些被绝大多数的编辑小组成员同意的代码被标注为“正确”。

把那些“正确”代码与原来人类编辑指定的代码进行比较是很有意义的。通过人工分配到报导的代码 88% 是正确的。然而，人类编辑是会犯错的：原来由人类编辑分配的代码总

计有17%是不正确的，如图8-4所示。

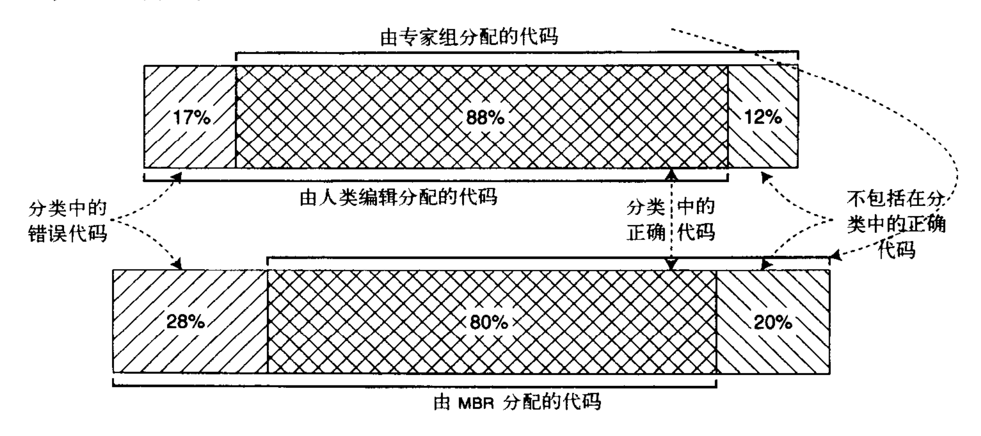


图8-4 由人类编辑和MBR给新闻报导分配代码的结果比较

MBR也没有做好。对于MBR，相应的百分率是80%和28%。即MBR分配的代码中有80%是正确的，但是，在分配的所有代码中，有28%的代码不正确。

最初的代码分配是由各种水平的编辑共同完成的，其中既包括新手，也包括了中等熟练程度及有经验的编辑。MBR系统的实际表现比新手好一些，与中等程度编辑的工作相当。MBR也在使用同样的由各种水平混杂的编辑分类出的报导作为训练集，因此，训练集编码方式不是完全一致。但令人惊讶的是，虽然给出的训练集有不一致的地方，MBR所给出的结果同编辑们做得几乎一样好。因为没有足够可用的这类报导作为训练集，训练集的代码是通过专家小组复审得来的，所以在这样的训练集中使用MBR，案例不能充分地做调查研究。

这个案例说明，MBR能解决不容易被其他方法解决的困难问题。大多数数据挖掘技术不能够处理文本数据，而且同时分配多个种类总是困难的。这个案例显示，藉由一些实验，MBR方法产生的结果能够与人类专家相媲美。对于如何测量MBR在评估文档分类或检索系统方面的表现，在后面“测量分配代码的有效性”部分中有进一步的讨论。这个案例达到这样的结果大约花费了两个人数月的努力（不包括相关性反馈引擎的开发时间）。但作为比较，像以专家系统为基础的那些其他的自动分类技术，对新闻报导分类大概需要许多人数年的努力才能达到相同的结果。

8.4 测量距离

假设你将要到一个小城镇去旅行，想要知道那儿的天气。如果有一份列出主要城市天气预报的报纸，通常你会做的就是找出那个小城镇附近大城市的天气。你可能把最靠近的城市当成该镇的天气，或者把三个最靠近的城市天气状况做出某种合并组合来预测其天气情况。这是使用MBR方法得出天气预报的实例，所使用的距离函数是在两个位置之间的地理距离。网上那些以邮政编码为区域提供天气预报的服务，采用的有可能就是这类原理。

8.4.1 什么是距离函数

距离是MBR测量相似性的一个手段。对于任何的真实距离测量，从点A到点B的距离，记为d（A，B），有四个主要的性质：

1）**定义明确**：在两个点之间的距离总是确定的，并且是一个非负实数，d（A，B）>0。

2）**同一性**：从一个点到它本身的距离总是零，因此 d（A，A）=0。

3）**可交换性**：方向不会造成任何差别，因此从 A 到 B 的距离与从 B 到 A 的距离是相同的：d（A，B）=d（B，A）。这一性质排除了单行道的情况。

4）**满足三角形不等式**：在从 A 到 B 的途中转道一个中间点 C，永远不会缩短距离，因此 d（A，B）≤d（A，C）+d（C，B）。

对于 MBR，所谓的点实际上是一条数据库记录。这个正规定义的距离是测量相似性的基础，当一些约束条件放松一点的时候，MBR 仍然可以相当好地完成工作。例如，新闻报导分类案例中的距离函数是不可交换的；即从一个新闻报导 A 到另外一个 B 的距离并不总是等同于从 B 到 A 的距离，但是相似性测量对分类目的仍然是有用的。

到底什么原因使这些性质对 MBR 有用？从有明确定义的距离这个概念可以给出一个推论：在数据库中的某处，每个记录都有一个邻居——MBR 需要找到邻居才能正常工作。同一性使距离遵从直觉的概念：给定记录的最相似记录就是原始记录本身。可交换性和三角形不等式使最近邻居成为局部的和表现良好的。在数据库中加入新记录不会使当前的记录靠得更近。相似性只是一次衡量两个记录的手段。

虽然通过测量距离找最近邻的方法工作良好，但最近邻的集合仍然有一些独特的性质。例如，一条记录 B 的最近邻居可能是 A，但是 A 可能有许多邻居比 B 靠得更近，如图 8-5 所示，但这种情形不会给 MBR 造成问题。

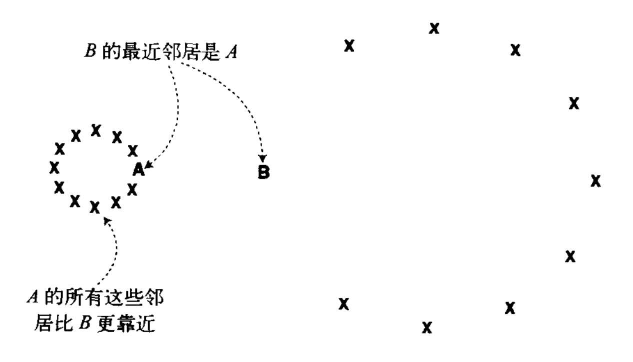


图 8-5　B 的最近邻居是 A，但是 A 有许多邻居比 B 靠得更近

测量分配代码的有效性：查全率（recall）和查准率（precision）

要确定一组分配代码或关键词是否恰当，查全率和查准率是两个有用的参数。例如，分配新闻报导代码的案例中，MBR 给新闻报导分配了许多代码。查全率和查准率能用来评估这些分配的好坏。

查全率可以回答这样的问题："MBR 把多少正确的代码分配到报导中？"它是 MBR 分配的正确代码数（经过编辑所验证的）与报导中的正确代码总数相除得到的比率。如果 MBR 分配到每个报导的所有代码都是正确的，查全率是 100%，因为正确的代码全部被分配，当然里面还包括许多其他无关代码。如果 MBR 不给任何报导分配代码，其查全率是 0%。

查准率回答的问题是："MBR 分配的所有代码中，有多少是正确的？"它是指 MBR 分配的正确代码数占 MBR 分配的代码总数的百分比。当 MBR 为报导分配的代码全部正确的时候，查准率是 100%；当 MBR 给每个报导分配所有代码的时候，查准率接近 0%。

单独给出的查全率或查准率本身并不能完整反映分类的好坏。当然最理想的情况是，我们想要100%的查全率和100%的查准率。通常，牺牲其中的一个可以提高另一个。例如，使用较多邻居可以增加查全率，但是减少查准率；反过来，提高阈值会增加查准率，但是减少查全率。表8-5给出了几个特定的案例，可以看出这些测量的一些相互影响。

表8-5 关于查全率和查准率关系的几个例子

MBR分配代码	正确代码	查全率	查准率
A，B，C，D	A，B，C，D	100%	100%
A，B	A，B，C，D	50%	100%
A，B，C，D，E，F，G，H	A，B，C，D	100%	50%
E，F	A，B，C，D	0%	0%
A，B，E，F	A，B，C，D	50%	50%

与校正过的正确代码集相比较，被编辑个人分配到报导的最初代码有83%的查全率，88%的查准率。对于MBR分配的代码，是80%查全率和72%的查准率。然而，表8-6给出的对所有分类的平均情况表明，MBR在某一些分类方面做得更好。

表8-6 按照代码种类给出的MBR查全率和查准率

种类	查全率	查准率
政府	85%	87%
工业界	91%	85%
市场区域	93%	91%
产品	69%	89%
区域	86%	64%
科目	72%	53%

由于分类而造成结果的差异意味着：可能没有给作为训练集的最初报导分配一致的代码。由MBR给出的结果最多只能与被选为训练集的样本的结果一样好。即使如此，MBR的表现几乎跟最有经验的编辑一样好。

8.4.2 每次每个字段只建立一个距离函数

从几何学的概念理解距离是很容易的，但是如何为具有许多不同类型不同字段的记录来定义距离呢？回答是：每次一个字段。下面来考虑如表8-7所示的样本记录。

表8-7 营销数据库中的五个客户情况

记录号	性别	年龄	薪金
1	女	27	$19 000
2	男	51	$64 000
3	男	52	$105 000
4	女	33	$55 000
5	男	45	$45 000

图8-6给出了一个三维的散点图。记录稍微有点复杂，是由两个数值字段和一个分类字

段构成的。这个实例显示了应该如何给每一个字段的距离函数下定义，然后把每个字段的距离函数组合在一起，变成一条记录的距离函数，从而给出两条记录之间的距离。

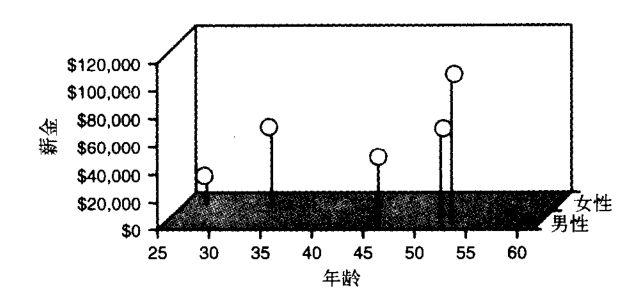


图 8-6 这个散点图以三个维度——年龄、薪金和性别图示了表 8-7 中的五条记录，而且表明标准距离对于最近邻居是一个好的度量

数值字段的四个最常用的距离函数是：

- 差的绝对数值：$|A-B|$
- 差的平方：$(A-B)^2$
- 归一化绝对值：$|A-B|$／（最大差值）
- 标准差的绝对值：|（A − 均值）／（标准差） − （B − 均值）／（标准差）|，等同于 |（$A-B$）／（标准差）|

归一化绝对值的优点是，它总是在 0 和 1 之间。在这个实例中，由于年龄比薪金的数值要小许多，对它们两个来说，归一化绝对值是一个好的选择——这样两者皆不会主导这条记录的距离函数（标准差也是一个好的选择）。对于年龄，距离矩阵如表 8-8 所示。

表 8-8 基于客户年龄的距离矩阵

	27	51	52	33	45
27	0.00	0.96	1.00	0.24	0.72
51	0.96	0.00	0.04	0.72	0.24
52	1.00	0.04	0.00	0.76	0.28
33	0.24	0.72	0.76	0.00	0.48
45	0.72	0.24	0.28	0.48	0.00

性别是一个分类数据的实例。最简单的距离函数是“恒等”函数，当性别相同的时候是 1，否则为 0：

d_{gender}（female，female）= 1

d_{gender}（female，male）= 0

d_{gender}（male，female）= 0

d_{gender}（male，male）= 1

所以非常简单。现在有三个字段的距离函数，需要组合为一个单个记录的距离函数。通常有三种最常用的方法：

- Manhattan 距离或代数和：

$$d_{sum}(A, B) = d_{gender}(A, B) + d_{age}(A, B) + d_{salary}(A, B)$$

- 归一化求和：d_{norm}（A，B）$=d_{sum}$（A，B）$/\max$（d_{sum}）
- 欧几里得几何距离：

$$d_{Euclid}(A, B)=\mathrm{sqrt}(d_{gender}(A, B)^2+d_{age}(A, B)^2+d_{salary}(A, B)^2)$$

表 8-9 显示了使用这三个函数得出的每一个点的最近邻居。

表 8-9 三个距离函数的最近邻居集合，按最近到最远排序

	d_{sum}	d_{norm}	d_{Euclid}
1	1，4，5，2，3	1，4，5，2，3	1，4，5，2，3
2	2，5，3，4，1	2，5，3，4，1	2，5，3，4，1
3	3，2，5，4，1	3，2，5，4，1	3，2，5，4，1
4	4，1，5，2，3	4，1，5，2，3	4，1，5，2，3
5	5，2，3，4，1	5，2，3，4，1	5，2，3，4，1

在这个例子中，最近邻居集合几乎是完全相同的，不管各个距离怎样组合。这是一个巧合，原因是五个记录恰好落在两个明确定义的簇中。簇之一是薪金比较低的年轻女性，另一个是薪金比较高的较年长的男性。这些簇暗示，相对于某一个字段，如果两个记录是彼此靠近的，那么它们在所有的字段都是相近的，所以在每个字段上距离的组合方式并不重要。通常的状况却不是这样的。

下面来考虑当一条新记录（见表 8-10）用于比较的时候，情况会怎样。

表 8-10 新客户的记录

记 录 号	性 别	年 龄	薪 金
新记录	女	45	$100 000

这条新记录不在其中任一个簇中。表 8-11 中列出了它与训练集的相应距离，以及它的邻居列表（从最近到最远排序）。

表 8-11 新客户的最近邻居集合

	1	2	3	4	5	邻 居
d_{sum}	1.662	1.659	1.338	1.003	1.640	4，3，5，2，1
d_{norm}	0.554	0.553	0.446	0.334	0.547	4，3，5，2，1
d_{Euclid}	0.781	1.052	1.251	0.494	1.000	4，1，5，2，3

现在的邻居集合就依赖于如何组合字段距离函数来求得记录的距离函数。事实上，使用求和和归一化得出的第二个最近邻居，是使用欧几里得几何得出的最远邻居，反之亦然。与求和和归一化度量相比，欧几里得几何度量更倾向于给出所有的字段都相对接近的邻居。它排斥记录 3，是因为性别不同，而且其距离最远（达 1.00 的距离）。因为性别是相同的，所以相应地它支持记录 1。注意 d_{sum}和 d_{norm}的邻居排序是相同的。归一化的距离定义保留了求和距离的排序——距离值的改变仅仅只是范围从 0 到 1 的改变。

求和、欧几里得几何以及归一化函数也能加上权重，这样每个字段对记录的距离函数可以贡献一个不同的量。通常所有的权重等于 1 时，MBR 给出好的结果。然而，有时可以利用权重来突出某个重要方面，比如一个疑似在类别划分方面有更大作用的特殊字段，就可以给它一个更高的权重。

8.4.3 其他数据类型的距离函数

一个5位数的美国邮政编码常常用一个简单数字代替。对于数字字段，任何一个默认的距离函数都有意义吗？不是！两个任意的邮政编码之差没有任何意义，更确切地说，是几乎没有意义。但毕竟一个邮政编码嵌入了位置信息，例如，前三个数字代表一个邮政区域，Manhattan 的邮政编码即是以“100”、“101”或“102”开头的。

此外，邮政编码从东到西有一个逐渐增加的模式。从0开始的编码在 New England 和 Puerto Rico；那些以9开始的编码在西海岸。这就暗示了这样一个距离函数：可通过察看邮政编码的前几位数估计地理距离远近。

- $d_{zip}(A, B)=0.0$　邮政编码相同的时候
- $d_{zip}(A, B)=0.1$　前三位数相同的时候（比如“20008”和“20015”）
- $d_{zip}(A, B)=0.5$　仅第一位数相同的时候（比如“95050”和“98125”）
- $d_{zip}(A, B)=1.0$　第一位数不相同的时候（比如，“02138”和“94704”）

当然，如果地理距离真的令人感兴趣，较好的方法是在表中查出一个邮政编码的经度和纬度，然后以同样的方式计算距离（关于这样的信息，在美国可以通过访问 www.census.gov 得到）。然而对于许多目的，地理位置的接近并不像其他衡量相似性的度量那么重要。10011 和 10031 两者都在 Manhattan，但是从市场营销的角度看，它们并没有更多相同的东西，因为一个是高消费阶层的市中心区域，而另一个是工人阶层的黑人住宅区；另一方面，02138 和 94704 处于方向完全相反的两个海岸，但可能同样地响应来自一个政治行动委员会的直接邮寄，因为它们分别对应于麻省的剑桥（Cambridge，MA）和加州的伯克利（Berkeley，CA）。

这只是距离选择怎样依赖于数据挖掘环境的一个实例。在第11章中会讲到另外一些测量距离和相似性的例子，在那里它们被应用于聚类。

8.4.4 当距离度量已经存在时

有一些情形，距离度量已经存在，但是难以发现，这样的情况通常以两种形式出现。其一，已经存在一个函数，通过适当改变可以作为 MBR 中使用的距离函数。新闻报导案例提供了一个好的实例，把已经存在的函数——相关性反馈得分——用作一个距离函数。

其二是，有一些字段虽然与距离不沾边，但也能够借用来为 MBR 服务。一个这类隐藏距离字段的实例是诱惑历史（solicitation history）：过去为一个特别的诱惑选择的两个客户是“相近的”，即使他们被选择的理由可能不再适用；设被选择的两个客户也是相近的，但是不如前者相近；而一个被选择的和一个没被选择的客户是遥远地分开的。这种度量的优势是它能融入以前的判断，即使当初判断的基础不再有效。另一方面，对那些在一开始就没包括在诱惑活动对象里的客户，它不会很好地起作用；所以某些类的中性权重必须加进来。

考虑初始被诱惑客户是否响应了最初的诱惑，能更进一步扩充这个函数，形成一个诱惑度量，如：

- $d_{solicitation}(A, B)=0$，　当 A 和 B 两者都响应诱惑
- $d_{solicitation}(A, B)=0.1$，　当 A 和 B 都被选择，但都未响应
- $d_{solicitation}(A, B)=0.2$，　当 A 和 B 只有一个被选择，但是两者的数据都有效

- $d_{solicitation}$（A，B）=0.3，　当 A 和 B 都被选择，但是只有一个响应时
- $d_{solicitation}$（A，B）=0.3，　当一个或两个不被考虑的时候
- $d_{solicitation}$（A，B）=1.0，　当一个被选择，而且另一个不被选择的时候

当然，这些特别的数值并非不可变更，它们只作为相似性测量指导，而且展示了以前的信息和响应历史该如何融入距离函数之内。

8.5 组合函数：向邻居求答案

距离函数用来决定哪条记录可以包含在邻居中。这一节将介绍通过组合不同邻居的数据做出预测的其他方法。在本章的开始，我们估计了 Tuxedo 镇的中值房租，采用的方法是取相似城镇的中值房租的平均。在那个实例中，平均就是组合函数。这一节将探究寻找邻居的其他方法。

8.5.1 基本的方法：民主

一个通常的组合函数是由 k 个最近邻居投票给出一个答案——在数据挖掘中发扬“民主”。当 MBR 用于分类的时候，每个邻居都会把票投给自己的类，从赞成每个类的票数比例可以估计新记录属于某相应类的可能性。当任务仅是分配一个单一类别的时候，新记录就属于有最多选票的那一个。当只有两个类的时候，所选的邻居数目应为奇数以避免出现平局。有一个经验法则，当有 c 个类的时候，至少要使用 $c+1$ 个邻居以保证某一分类有一个相对多数。

在表 8-12 中，是先前提到的五个测试记录，针对客户是否已经变得不活跃增加了一个标志。

表 8-12　具有流失历史的客户

记录号	性别	年龄	薪金	不活跃
1	女	27	$19 000	No
2	男	51	$64 000	Yes
3	男	52	$105 000	Yes
4	女	33	$55 000	Yes
5	男	45	$45 000	No
新记录	女	45	$100 000	?

对于这个实例，其中的三个客户已经变得不活跃，另外两个依然活跃，这是一个几乎平衡的训练集。出于说明的目的，如果使用 k 的不同数值作为两个距离函数 d_{Euclid}和 d_{norm}（见表 8-13），让我们试着确定新的记录是活跃的还是不活跃的。

表 8-13　使用 MBR 确定新客户是否会不活跃

	邻居	邻居流失	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
d_{sum}	4，3，5，2，1	Y，Y，N，Y，N	yes	yes	yes	yes	yes
d_{Euclid}	4，1，5，2，3	Y，N，N，Y，Y	yes	?	no	?	yes

问号表示由于邻居打成平手而没有给出明确预测的情况。可以看出，k 的不同数值确实

影响到分类状况。这说明使用一致的邻居百分率可以提供对置信度水平的预测（表 8-14）。

表 8-14 带置信度的客户流失预测

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
d_{sum}	yes，100%	yes，100%	yes，67%	yes，75%	yes，60%
d_{Euclid}	yes，100%	yes，50%	no，67%	yes，50%	yes，60%

当有两个以上种类时，同样可以用置信度水平进行判断。然而，有较多类的情况下，很可能没有单个类会有绝对多数票。MBR（和大多数数据挖掘方法一样）的主要假设之一是，训练集能为预言性目的提供充分的信息。如果新记录的邻居们总是不能给出新记录类别的明显选择，那么数据也许没有包含必要的信息，维度的选择甚至可能训练集的选择都需要重新评估。通过测量 MBR 在测试集上的有效性，你能判断训练集是否有足够数目的样本。

警告：MBR 能达到的最好水平是和它使用的训练集一样。为了测量训练集是否有效，可以在测试集中使用 2、3 和 4 个邻居，观察它预测的结果，如果结果是不确定的或不准确的，那么原因可能是，训练集不够大，或者维度和距离度量选择不当。

8.5.2 加权投票

加权投票（weighted voting）与前一节中的投票类似，只不过邻居并不是完全平等的——更像大小不同的股东们的民主，不是一人一票。选票的比重与距新记录的距离成反比关系，因此，近的邻居比远的邻居有更高的选票比重。为了避免距离可能是 0 的问题，通常在取倒数以前，把距离加 1，增加 1 也使得所有的选票权重值在 0 和 1 之间。

表 8-15 把加权投票应用到前述实例中。“是的，客户将会不活跃”的选票是第一，那些“不，这是好客户”的选票排第二。

表 8-15 使用加权投票预测客户流失

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
d_{sum}	0.749～0	1.441～0	1.441～0.647	2.085～0.647	2.085～1.290
d_{Euclid}	0.669～0	0.669～0.562	0.669～1.062	1.157～1.062	1.601～1.062

加权投票已经引进充足的变化避免出现平局，置信度可以由赢得选票与总选票的比率（表 8-16）进行计算。

表 8-16 加权投票给出的置信度

	1	2	3	4	5
d_{sum}	yes，100%	yes，100%	yes，69%	yes，76%	yes，62%
d_{Euclid}	yes，100%	yes，54%	no，61%	yes，52%	yes，60%

在这个例子中，为选票加权对结果和置信度只有小的影响。而当一些邻居距另外的邻居相当远的时候，加权的影响是最大的。

如果用加权距离的平均值来代替邻近值的简单平均，加权也可以用于估计（estimation）。这种方法可以用于协同过滤体系，正如下一节将要描述的。

8.6 协同过滤：可以做出推荐的最近邻方法

本书的作者没有一个人认为自己是乡村音乐迷，但是其中的一个却自豪地拥有演唱者亲笔签名的 Dixie Chicks 乐队的早期 CD。Chicks 乐队，那时还不是一个响亮的乐队，当时正在一个当地酒吧进行表演，来自德克萨斯的一些认识他们的朋友作了热情的推荐。表演真的令人难忘，尤其是 Martie Erwin 完美无瑕的蓝草音乐小提琴，她妹妹 Emily 演奏的各种令人眼花缭乱的乐器（大多数是弦乐器，但是有些不是），和 Laura Lynch 那富有魅力的声音（她也弹一把电贝司）。在演出间隙，她们曾亲笔签名售卖自制的 CD，我们更喜欢这盘 CD，尽管后来的 CD 为她们赢得了葛莱美奖。这个例子与最近邻居技术有什么关系吗？对了，这就是一个应用协同过滤的实例。来自信赖朋友的推荐将会导致一个人尝试某事，如果没有这个推荐的话，他根本不可能去试。

协同过滤是一个基于存储的推理的变形，特别适合于提供个性化推荐方面的应用。协同过滤系统从人们的喜好历史开始。距离函数基于喜好重叠确定相似性——喜欢同样事物的人是相近的。此外，选票有一个距离权重，因此，较近的邻居选票在推荐中占的比重更大。换句话说，它是使用一个已选出的同年龄组的相似口味为判据，发现音乐、书籍、酒或其他东西是否适合某个人的技术。这一方法也被称作*社会信息过滤*（social information filtering）。

协同过滤是利用口头语言确定人们是否喜欢某事的自动化过程。知道许多人喜欢某件事并不够，哪些人喜欢它也是重要的。每个人对于一些推荐的评价总可能比另外一些人更高。过去曾经做过正确推荐的亲密朋友的推荐可能足以让你去看一场新电影，即使它属于一个你通常不喜欢的流派。另一方面，一个朋友认为“Ace Ventura：Pet Detective”是他看过的最好笑电影，热情地推荐给你，但你可能不去看这样一部你原来可能想看的电影。

使用自动协同过滤系统为一个新客户准备推荐有三个步骤：

1）让新客户估价一些挑选出的项目，如电影、歌曲或餐馆，建立一个客户简档。

2）使用某种相似性度量，比较新客户的简档与其他客户的简档。

3）按照简档的相似性把客户分级，利用客户分级组合可以预测新客户会把他（或她）没有分级的项目分到哪一个级。

下面几节将更详细地介绍上述步骤。

8.6.1 建立简档

协同过滤的一个挑战是，时常有较多的项目需要分级，这比任何人可能遇到的或者愿意做的都要多。简档通常是不足的，这意味着用作推荐的使用者喜好之间有很少重叠。在将要分级的大量项目中，可以把使用者简档看做一个矢量，简档中的每个项目对应于矢量中的一个分量（element）。矢量的每个分量代表简档所有者对某个项目的分级情况，项目按照从 −5到 5 的等级划分，0 表示中立，空表示没有评价意见。

如果矢量中包含数千或数万个分量，而且每个客户自己决定分级哪一些项目，那么最后很可能任意两个客户的简档都仅有极少数重叠。另一方面，让客户对一个特别的子集分级，可能错过重要的信息，因为较模糊项目的分级可能侧重于客户自身，而不是对通常项目的分级状况，比如对披头士乐队的喜爱比对 Mose Allison 的喜爱显示的信息更少。

一个合理的方法是让新客户分级 20 个左右最常分级的项目（当然这个列表可以随时间

改变），然后请他们自由地给尽可能多的额外项目分级。

8.6.2 比较简档

建立客户简档以后，下一步工作就是测量它与其他简档的距离。最直接的方法可能是视简档矢量为几何学的点，计算它们之间的欧几里得几何距离，当然，人们还尝试了许多其他的距离度量方法。为了与最后的结果一致，有时会给一个正面分级的用户较高的权重，尤其是在多数用户给多数项目负面分级的时候更是如此；当然，另外一些人把统计关联测试应用到矢量分级中。

8.6.3 做出预测

最后一步是使用一些近邻简档的组合，为客户尚未分级的项目给出一个估计的分级。一种方法是取加权平均，所用的权重与距离成反比。图 8-7 所示实例说明，以邻居 Simon 和 Amelia 的意见为基础，如何估计 Nathaniel 给电影“Planet of the Apes”的分级。

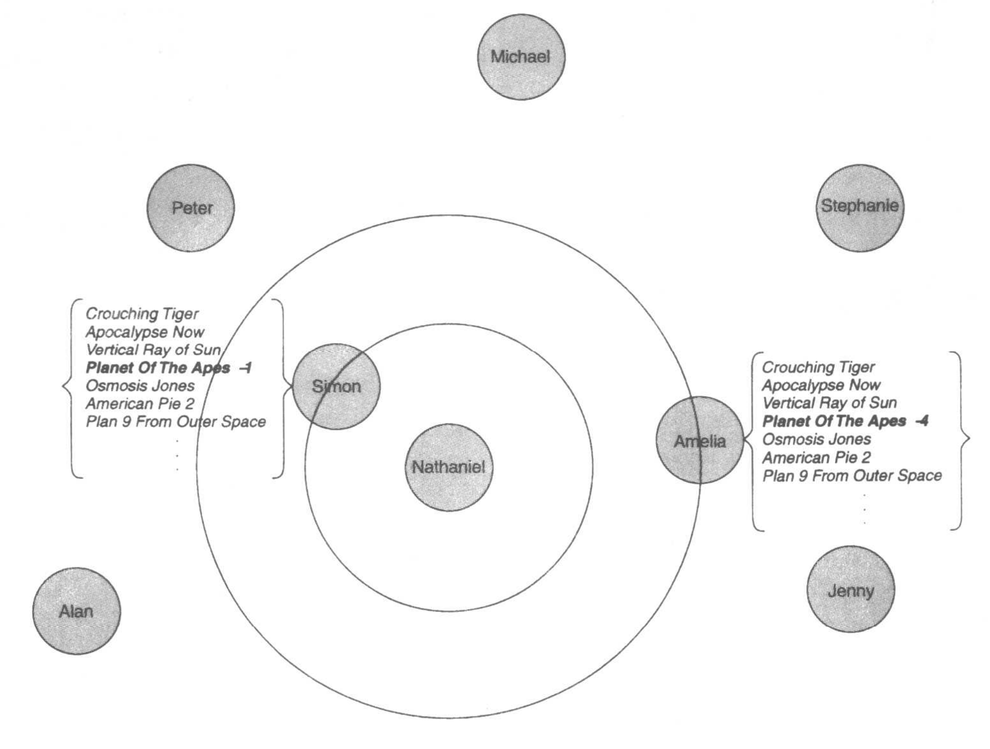


图 8-7 预测给电影“Planet of the Apes”的等级分是 -2.66

距离是 2 的 Simon 给予该电影一个 -1 的分级，距离是 4 的 Amelia 给予该电影一个 -4 的分级。在这次投票中，没有其他人的简档与 Nathaniel 的简档更接近。因为 Amelia 的距离是 Simon 的两倍，她的选票权重只及 Simon 的一半，按照距离权重为 Nathaniel 的分级估计是：

$$(1/2\ (-1)\ +1/4\ (-4))\ /\ (1/2+1/4)\ =-1.5/0.75=-2$$

好的协同过滤系统给使用者提供机会来适当地评论预测，然后调整简档。在这个实例中，尽管 Nathaniel 有将不喜欢电影“Planet of the Apes”的预测，但如果他真的去租借了这

个电影的录像带，他就会有一个自己的实际分级。如果他真的喜欢这部电影而且给它一个4的分级，他的新简档邻居将会有稍微不同的改变，Simon 和 Amelia 的意见将会在对 Nathaniel 的下一次推荐中占比较少的权重。

8.7 小结

基于存储的推理是一项有力的数据挖掘技术，能用来解释各式各样的数据挖掘问题，包括分类和估计。其他的数据挖掘技术，一般需要先使用一个有预分类数据的训练集产生一个模型，然后再抛开训练集工作；而对于 MBR 不是这样，它的训练集实际上就是模型本身。

选择正确的训练集大概是 MBR 最重要的一步。训练集需要包括覆盖所有可能类别的样本。这说明，为了产生一个对所有的类都有大约相同数目实例的平衡训练集，对于罕见分类，需要包含一些数目不成比例的罕见类别实例以丰富该训练集。如果训练集只包括差的客户实例，给出的预测结果可能是：所有的客户都是差的。一般来说，训练集如果没有数十万或数百万个样本的话，至少应该有数千个样本。

MBR 是一种寻找 k 个最近邻居的方法，决定邻居的远近需要一个距离函数。有许多方法可以测量两个记录之间的距离，仔细选择适当的距离函数对 MBR 的使用是非常关键的一步。本章引入了一个方法：通过为每个字段建立一个距离函数而且使它归一化，产生一个总的距离函数。这个归一化的字段距离可以通过欧几里得几何的方式组合，或者求和来产生一个 Manhattan 距离。

当使用欧几里得几何方法时，任何一个字段中出现大的差别都足以导致要考察的两个记录远远分开。Manhattan 方法更宽容一些——在一个字段上的大的差别可以容易地被其他字段上的接近所抵消。通过将模型集应用于所有候选距离函数以找到给出较好结果的函数，验证集可以为一个给定的模型找出最佳距离函数。邻居的正确选择有时需要调整距离函数，以使一些字段比另外一些更有利，这可以通过将权重引入距离函数来容易地完成。

下一个问题是选择邻居数目。再次利用验证集找出不同数目的邻居可以帮助确定邻居的最佳数目。邻居的数目实际上没有一个确切值，因为该数目依赖数据的分布状况，而且与被解决的问题密切相关。

基本的加权投票式组合函数对于分类数据工作良好，所使用的权重与距离成反比。用于估计数字型数值的类似运算是取一个加权平均。

基于存储的推理方法的一个很好应用就是做出推荐。协同过滤也是一个做出推荐的方法，它使用距离函数来比较两个用户分级列表，把具有相似口味的人归为一组。给一个新人作推荐是通过加权平均他或她最近邻居的分级来计算的。

第 9 章　购物篮分析和关联规则

为传达购物篮分析（market basket analysis）的基本思想，我们将从图 9-1 所示商店手推车的情形开始，手推车中装满了某人到一家超市购物时购买的各种产品。这个购物篮包含了橙汁、香蕉、软饮料、擦窗器和清洁剂。购物篮告诉我们客户会同时购买什么。所有客户的完整购物列表提供很多信息，描述了零售业的最重要信息——客户在购买什么商品以及何时购买。

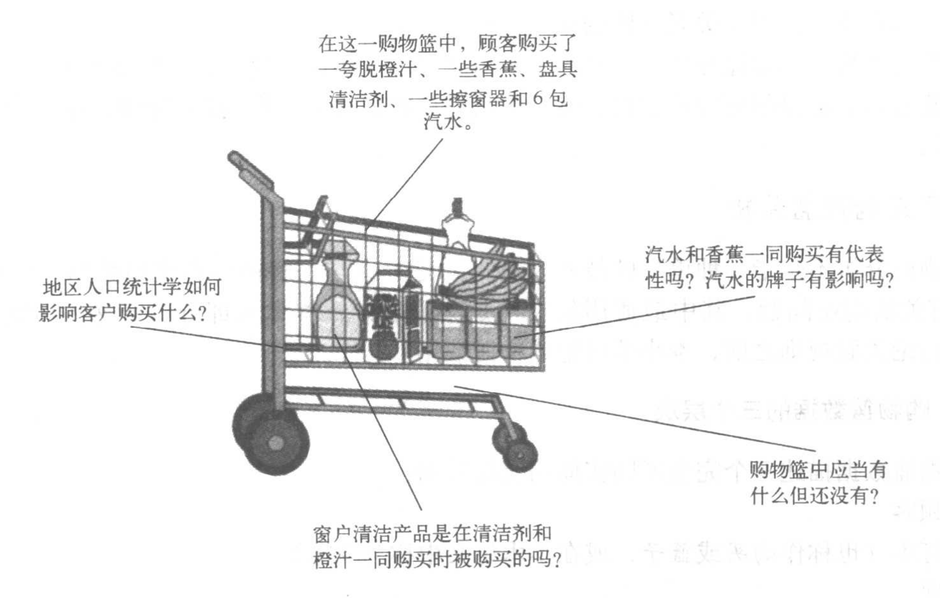


图 9-1　购物篮分析不仅有助于理解客户，也有助于理解他们一同购买的项

每个客户都会购买不同产品，其购买数量不同，购买时间不同。购物篮分析使用关于客户购买物品的信息，来深入了解他们是谁和他们为什么做出某种购买。购物篮分析通过告诉我们哪些产品倾向于被一起购买和哪些产品最有必要促销来提供对产品的深入了解。这些信息是可操作的：它能建议新的铺货规划，确定哪些产品上架，指导何时发优惠券等。如果这一数据能够通过忠诚卡或网站注册与个别客户结合起来，就变得更有价值了。

与购物篮分析最密切的数据挖掘技术是关联规则（association rule）的自动生成。关联规则代表没有特定目标的数据中的模式。同样，它们是非定向数据挖掘（undirected data mining）的实例，模式是否有意义由人类作出解释。

关联规则最初用于描述诸如什么产品被同时购买的销售点数据（point-of-sale）。尽管它最初是用于分析销售点交易，但也能够应用于零售业之外去发现其他类型的“篮子”之间的关系。一些可能的应用例子是：

- 用信用卡消费的项，诸如租车和旅馆房间，提供对客户可能购买的下一产品的深入了解。
- 移动通信客户购买的可选服务（呼叫等待、呼叫转移、数字用户线路［Digital Subscriber

Line，DSL]、快速呼叫等)，帮助确定如何把这些服务捆绑到一起，以获得最大收益。

- 零售客户使用的金融服务（现金销售账户、现金付款机［Cash Dispenser，CD]、投资服务、购车贷款等)，识别客户可能想要的其他服务。
- 不寻常的保险索赔（insurance claim）组合可能是欺诈的征兆，能够激发深入的调查。
- 医学病史能基于治疗的特定组合指示可能的并发症。

关联规则常常无法实践期望。例如在我们的经验中，它们对在诸如小额银行业务等领域建立交叉销售（cross-selling）模型并不是一个好的选择，因为规则以描述前一营销促销活动而告终。同样，在小额银行业务中，客户以支票账户（checking account）开始，然后成为储蓄存款账户（savings account）也很有代表性。产品之间的区别直到客户拥有更多的产品才会发现。本章不仅涵盖了关联规则的用途也包含其缺陷。

本章首先概述购物篮分析，包括对于不需要关联规则的购物篮数据的更基本分析。此后转向关联规则，解释如何得到它们，然后继续讨论扩展关联规则使之包括购物篮分析其他方面的方法。

9.1 定义购物篮分析

购物篮分析不是指一种单一的技术，它指的是一组与了解销售点交易数据（transaction data）有关的商业问题，其中最通用的技术是关联规则，本章大部分内容深入研究这个课题。在讨论关联规则之前，本小节讨论购物篮数据。

9.1.1 购物篮数据的三个层次

购物篮数据描述三个完全不同实体的交易数据：

- 顾客
- 订单（也称作购买或篮子，或在学术论文中称为项集）
- 项

在关系数据库中，购物篮数据的数据结构看上去与图9-2类似，这一数据结构包括四个重要的实体。

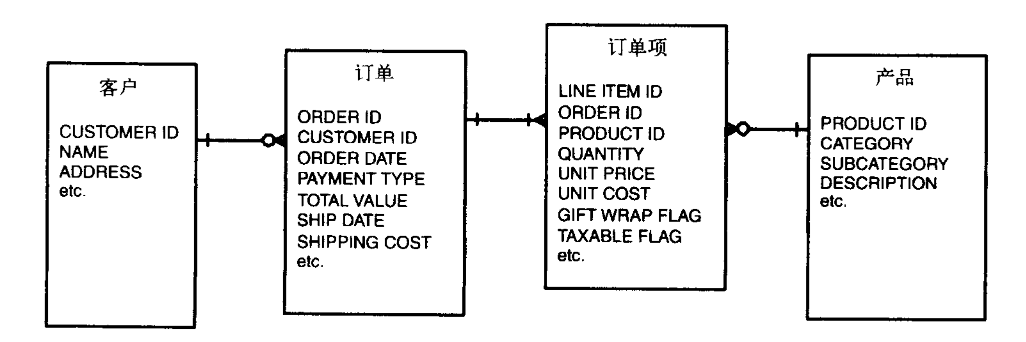


图9-2 代表交易层次购物篮数据的数据模型通常有三个表，一个关于客户，一个关于订单，一个关于订单项

订单是购物篮数据的基本数据结构。一个订单代表客户的单个购买事件。这可能对应于客户在网站上订购若干产品，或客户购买一篮子杂货，或客户从一个目录中购买了若干项。

这包括购买的总量、总金额、额外的运输费用、支付类型，和与该交易相关的任何其他数据。有时该交易被赋予一个惟一标识符。有时该惟一标识符需要根据其他数据进行调整。在一个例子中，我们需要组合四个字段得到代表在商店购物的一个标识符——客户支付的时间戳、连锁店代码（ID）、商店代码和街巷代码。

订单中的单个项分别表示为*订单项*，包括该项的支付价格、单项产品的数量、是否应当收取税款，或许还有成本（能够用于计算利润）。项表当中通常还有一个到*产品参照表*的链接，对每一产品提供更多的描述信息。这一描述信息包括产品分层和其他可能证明对分析有价值的信息。

*客户表*是一个可选的表格，当客户能够被识别时应当是可用的，例如，在一个需要注册的网站上或当交易过程中使用亲情卡的时候。客户表可能含有不同的令人感兴趣的字段，其中最有吸引力的部分是 ID 本身，因为这能把交易和时间结合起来。

随时间跟踪客户使得确定一些情况成为可能，例如，哪个食品购买者是“烘烤食品自助”者——这是面粉制造商和出售预先包装好的混合蛋糕粉制造商都非常感兴趣的。可以通过客户购买面粉、发酵粉和类似配料的频率，这种购买占客户总消费的比例，以及对预先包装的混合物和即食甜点缺乏兴趣等信息来识别这些客户。当然，这样的配料可能是在不同时间、以不同数量购买的，需要配合时间把多种交易结合起来。

购物篮数据的所有三个层次都是重要的。例如，要了解订单，有一些基本的度量：

- 每位客户的平均订单数是多少？
- 每一订单的特定项的平均数目是多少？
- 每一订单平均有多少项？
- 对于给定的产品，曾经购买过该产品的客户比例是多少？
- 对于给定的产品，包含该项的每个客户的平均订单数是多少？
- 对于给定的产品，当该产品被购买时一个订单中的平均购买量是多少？

这些度量给出了对该商务的广泛理解。在一些情况下，很少有重复的客户，因此每位客户的订单比例接近 1，这表明了一个商业机会，即增加每位客户的购买量。或者，每个订单产品的数量可能接近 1，表明在下订单的过程中进行交叉销售的机会。

把这些度量相互比较可能是有用的。我们已经发现：订单数常常是划分客户的有用方式，好的客户明显比不好的客户订购频率高。图 9-3 试图对于购买一项以上产品的客户通过客户关系的深度（订单数目）观察客户关系的宽度（曾购买的特定项的数目）。这一数据来源于一家小的专卖店。最大的泡泡显示购买两种产品的许多客户是在同一时间购买的。同样有一个令人惊讶的大泡泡，显示相当数量的客户用两个订单购买同样的产品。较好的客户——至少是那些多次回头的客户——倾向于购买更多样性的产品。然而，他们中有一些是回来购买他们第一次买到的同样东西。零售商如何鼓励客户回头买更多的不同产品？购物篮分析不能回答这个问题，但它至少能启发我们提出这个问题，并可能提供或许有所帮助的线索。

9.1.2 订单特征

客户购买行为有另外的重要特征。例如，平均订单大小随时间和地域不同而不同——追踪这些信息有助于了解在商业环境中的变化。这类信息在报告系统中常常是可用的，因为它很容易汇总。

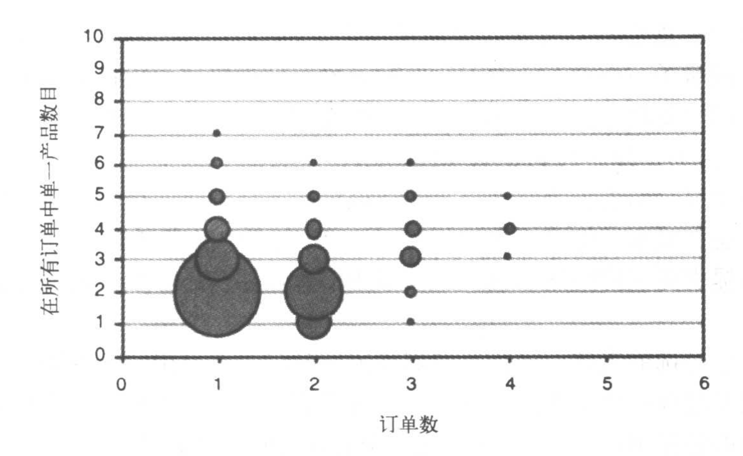


图 9-3 这一泡泡图通过关系的深度显示了客户关系的宽度

然而，有一些信息可能需要从交易层面的数据中收集。图 9-4 对另一个零售商按订单大小和支付信用卡——维萨信用卡、万事达信用卡或美国万国宝通银行卡——划分不同的交易。首先注意到的是不管使用哪种信用卡，订单越大，平均购买额越大，这是可靠的。同样，美国万国宝通银行卡这类信用卡的使用一向与较大的订单相关联——这是一个关于这些客户的重要发现。

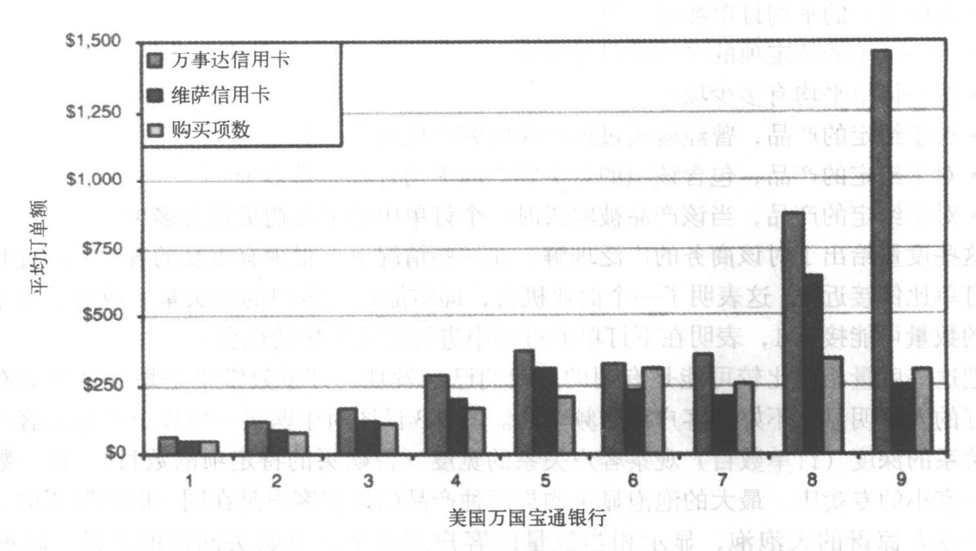


图 9-4 本图显示了对于一个特定零售商，以订单项数为基础，按信用卡消费的平均金额

对于网上购买和邮寄订单交易，在销售点也可能收集到额外的信息：

• 该订单使用礼品包装吗？
• 该订单转向与账单相同的地址吗？
• 购买者接受还是拒绝特定的交叉销售服务？

当然，在销售点收集信息和它对于分析可用是两回事。然而，礼品赠送和对交叉销售服务的响应对于了解客户是两件很有用的事情。用这一信息发现模式需要在第一现场（在呼叫中心或通过联机界面）收集信息，然后把它移入数据挖掘环境。

9.1.3 项流行性

什么是最流行的项？这是一个通过观察存货曲线能够回答的问题，不必使用交易层面的数据就能够生成。然而，知道单个项的销售仅仅是个起点，还有一些相关的问题：

- 在只有一个项的订单中发现的最普遍项是什么？
- 在有多个项的订单中发现的最普遍项是什么？
- 在重复购买的客户中发现的最普遍项是什么？
- 特定项的流行性随时间如何变化？
- 地域不同，一个项的流行性如何变化？

前三个问题特别值得关注，因为它们可能对客户关系的成长提出一些想法。关联规则能够对这些问题提供答案，特别是当与虚拟项（virtual item）一起使用以表示一个客户的订单大小或订单数的时候。

后两个问题提出了时间维度和地域维度，这对于购物篮分析的应用是非常重要的。不同的产品在不同地域有不同的吸引力——这是零售商非常熟悉的事。通过引入代表地域和季节的虚拟项，使用关联规则开始了解这些方面也是可能的。

提示：时间和地域是购物篮数据的最重要的两个属性，因为它们常常指向在特定销售时间的确切交易条件。

9.1.4 跟踪市场干预

正如在第 5 章讨论的，观察单个产品随时间的变化能对该产品正在发生的事情提供深入了解。把随时间而变化的市场干预连同产品销售一起考虑，如图 9-5 所示，则可能看到干预的效果。该图显示了一个特定产品的销售曲线。在干预之前，销售悬停在每周 50 个单位，而干预之后，销售峰值大约是该数量的 7 或 8 倍，尽管在六七周的时间内呈逐渐下滑之势。使用这种图，有可能测量该市场营销工作的响应情况。

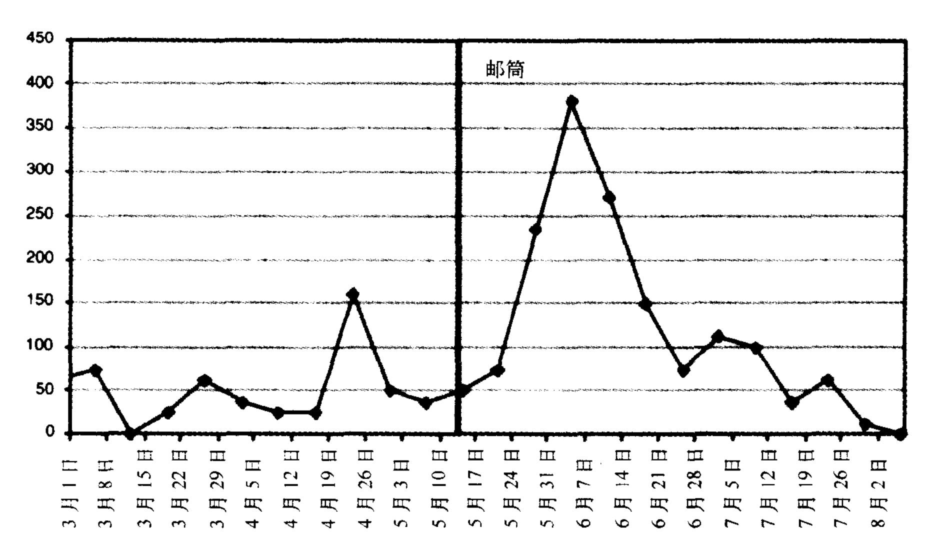


图 9-5　把市场干预和产品销售显示在同一图表中，可能看到市场营销工作的效果

这种分析不要求观察单个购物篮——每天或每周的产品销售合计就足够了。然而，它确实需要知道市场营销干预（marketing intervention）在何时发生的——有时要得到这样一个日程表也是巨大的挑战。这种图表能够回答的一个问题是干预的效果，但要回答这个问题，则必须确定额外的销售是逐渐增长的，还是由那些在晚些时候无论如何都将购买该产品的客户做出的。

购物篮数据能够回答这一问题，除了在干预之后观察销售量之外，也可以观察包含干预项的篮子的数目。如果客户数目不是在增加，就有证据说明现有的客户只不过是在该项上低价囤积。

一个相关的问题是，打折是否导致了其他产品的额外销售。通过发现促销阶段促销产品的组合，关联规则能帮助回答这个问题。类似地，我们可能想要知道在干预之后订单的平均额是增加了还是减少了，这些都是这类问题的示例，其中更详细的交易层数据很重要。

9.1.5 按用途聚类产品

也许最让人关注的一个问题是什么样的产品组合经常一起出现。这种产品组合对于向客户做出推荐是非常有用的——那些购买了某些产品的客户可能对其余的部分产品感兴趣（第8章更详细地讨论了产品推荐）。在单个产品分层上，关联规则在这一方面提供了一些答案。尤其是，这一数据挖掘技术可用于确定哪种或者哪些产品可以建议与其他特定产品同时购买。

有时我们想要发现那些比关联规则提供的聚类更大的聚类，关联规则中的任何规则提供的聚类只包括少数一些项。在第11章中描述的标准聚类技术也可用于购物篮数据。在这种情况下，数据需要转轴，如图9-6所示，以便每行代表一个订单或顾客，对购买的每一件产品有标记或计数器。可惜的是，常常有几千种不同的产品，为减少列的数目，这种转换可以在分类层面上进行，而不是在单个产品层面。

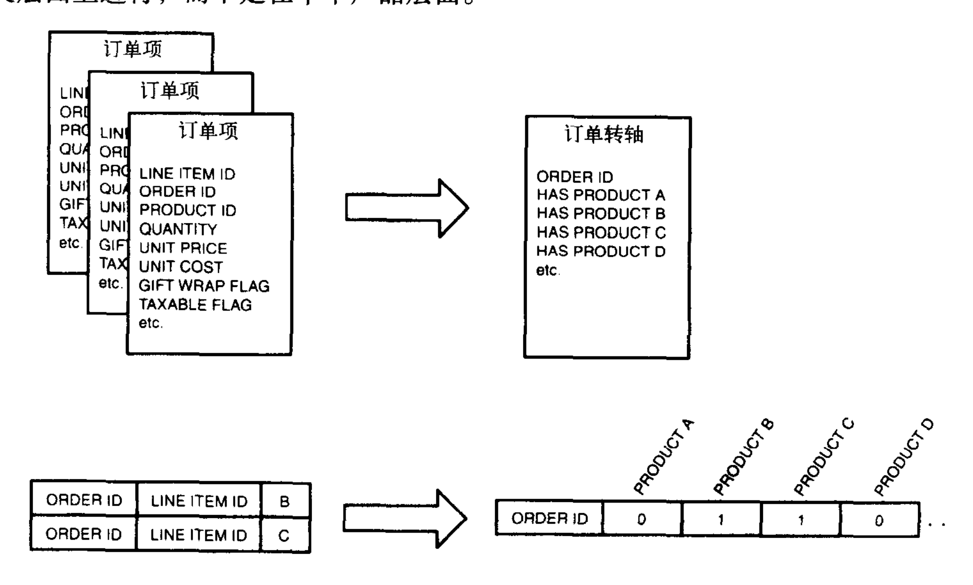


图9-6 转轴购物篮数据使得运行聚类算法有可能发现特别的产品群组

通常有大量关于产品的可用信息，除产品分层以外，还包括服装颜色、食品是否是低卡路里的、海报是否有背景等。这种描述提供了很多信息，能够引出特别有用的问题：

- 节食产品趋于一同销售吗？
- 客户在同一时间购买同色系的服装吗？
- 购买海报上产品的客户也买其他产品吗？

有能力回答这样的问题常常比试图聚类产品更有用，因为这种指导性信息常常直接导致市场营销行动。

9.2 关联规则

对关联规则的一个要求是，结果要有透明度和可用性，这些结果是以产品群组规则的形式表示的。关联规则要求数据挖掘者有较好的直觉，因为它表述了现实的产品和服务是如何组合到一起的。这样的规则是很清楚的："*如果一位客户购买了三方通话，则那位客户也将购买呼叫等待*"。甚至更好一些，它可能建议一个特殊的做法，诸如把呼叫等待和三方通话捆绑进单个服务包。

关联规则容易理解，但它们并不总是有用的。下列三个规则是从真实数据中生成的真实规则的例子：

- 购买芭比娃娃的沃尔玛（Wal-Mart）客户有60％的可能也购买三种类型的糖块之一。
- 签购维修保养协议的客户很可能购买大型家用电器。
- 当一个新的五金店开张时，最常卖的项之一是抽水马桶清洁器。

后两个例子是我们从数据中看到的真实例子，第一个是在1997年9月8日的《财富（福布斯)》杂志上引用的一个例子。这三个例子说明了关联规则产生的规则的三种普遍类型：*可操作的、平凡的和费解的*。除了这类规则之外，后面"著名的规则"部分谈到了另一种分类。

9.2.1 可操作的规则

有用的规则包含高质量、可操作的信息。一旦模式被发现，它常常不难证明，而如果是谎言的话，可以让我们有更深的了解并采取相应行动。芭比娃娃喜欢巧克力块胜过其他形式的食物不像是谎言。设想一个家庭出来购物，为的是为小苏茜的朋友埃米莉买一个生日礼物，芭比娃娃是理想的礼物。在收银台，小雅各布开始哭了。他也想要一点东西——块糖正合需要。也许埃米莉有个弟弟，不能不给他赠送一个小礼物。糖块也许是给妈妈买的，因为购买芭比娃娃是一个累人的活，因此妈妈需要补充一些能量。这些情景都说明糖块是附加于芭比娃娃之上的刺激购物。

沃尔玛是否能够利用这一信息还不清楚，但这一规则可能暗示更有效的产品布局，诸如确保客户从芭比娃娃购物区回来时必须走过糖果通道。它也可能暗示产品组合销售以及将糖块和玩偶组合在一起的促销活动；也可能暗示特别的产品广告方式。因为该规则容易理解，所以它暗示出似是而非的原因和可能的干预方式。

9.2.2 平凡的规则

平凡的结果早已被熟悉商业的任何一个人所知晓。第二个例子（"签购维修保养协议的客户很可能购买大型家用电器"）是一个平凡规则（trivial rule）的例子。实际上，客户通常会同时购买维修保养协议和大型家用电器。他们为什么还要另外购买维修保养协议？这二者是在一起做广告，并且很少独立出售（尽管当独立出售时，客户通常是购买大型家用电器而

不带协议，而不是购买协议却不带家用电器）。但这一规则是在分析 Sears 的数十万计的销售点交易之后发现的，尽管正确并在数据中得到很好的支持，但仍然是无用的。类似的结果很多：买木钉的人们也买钉子，买油漆的客户也买油漆刷，油和滤油器通常一起购买，牛肉饼和汉堡包、木炭和液体点火器，等等。

微妙的问题可能落入相同的范畴。例如，在本地电话服务上购买三方通话的人们几乎总是购买呼叫等待，这类看上去有趣的结果是过去交易计划和产品捆绑的结果。在电话服务选项中，三方通话通常与呼叫等待捆绑在一起，因此难以分别订购。在这一情况下，该分析不产生可操作的结果，而产生早已遵照行事的结果。购物篮分析特别容易再现前一营销活动的成功，这对于任何数据挖掘技术都是危险的，产生这一现象的原因是由于它依赖于未汇总销售点数据——这恰恰是用于定义活动成功的数据。从购物篮分析中得到的结果也许只不过是测定前一市场营销活动的成功。

平凡规则确实有一个用处，尽管它不是直接的数据挖掘应用。当一个规则应当在该时间100%出现，然而它却没有这样做，这种情形可以提供关于数据质量的许多信息。换句话说，不遵循平凡规则的例外情况指出了商业运作、数据收集和处理等可能需要进一步改进的方面。

9.2.3 费解的规则

费解的规则（inexplicable rule）似乎没法解释，并且不给出行动过程。第三种模式（“当一个新的五金店开张时，最常卖的项之一是抽水马桶清洁器”）是令人迷惑的，它用一个新的事实吸引我们，但提供了一些并不能深入了解消费者行为或商品或者暗示更多对策的信息。在这一案例中，一家大型五金公司针对新店开张发现了这一模式，但是没能找出如何从中受益。在商店开张期间许多项在打折销售，但是抽水马桶清洁器的销量很突出。更多的调查可能给出一些解释：抽水马桶清洁器比其他产品的折扣更低吗？它们是否在商店开张时被放置在人流密集的区域但在其他时间不易看见？从其他连锁商店的情况看，这个结果反常吗？在其他时间它们是难以发现的吗？无论什么原因，只用购物篮数据作进一步分析能够给出一个可信的解释，这一点值得怀疑。

警告：当应用购物篮分析时，许多结果常常是平凡的或费解的。平凡规则再现商业常识，浪费了利用高级分析技术的努力。费解的规则是数据中的偶然事件，是不可操作的。

著名的规则：啤酒和尿布

或许谈得最多的曾经“发现”的关联规则是啤酒和尿布之间的关联。这是在20世纪80年代末或90年代初期很有名的故事，那时计算机刚刚变得足够有能力分析大量数据。故事发生在美国中西部的某地，一个零售商在分析销售点数据以发现有益的模式。

你瞧！潜藏在所有交易数据中的是啤酒和尿布一起销售这一事实，这立即使得有营销头脑的人兴奋起来，他们想断定到底发生了什么。头脑中一闪念可能会做出如下解释：喝啤酒的人不想中断他们欣赏电视体育运动节目，因此他们买了尿布以减少到盥洗室的次数。不，情况并不是这样的。更可能的情况是，有小孩子的家庭准备周末休假，尿布是给孩子的，啤酒是给爸爸的。爸爸大概知道，在他喝完几瓶啤酒之后，妈妈将给孩子换尿布。

这是个有说服力的故事。撇开分析论，零售商能够用这一信息做什么呢？有两种相左的观点：一种说把啤酒和尿布紧靠着放在一起，于是当顾客购买了其中一个时，会记得去买另一个。

另一种说把它们放置得尽可能远，于是客户必须走过尽可能多的货架，从而有机会购买更多的项。商店还能把高利润的尿布和啤酒放得近些，尽管把婴儿产品和酒混合摆放大概是不适宜的。

这个故事是如此有影响力，作者注意到，至少四个公司在使用这个假说——IBM、Tandem（现在是 HP 的一部分）、Oracle 和 NCR Teradata。这个真实的故事是 1998 年 4 月 6 日在《财富（福布斯）》杂志的一篇文章中披露的，名为“啤酒－尿布综合症”。

所披露的故事至今依然是一种启示。显然，啤酒和尿布的销售基于存货清单被认为是相关的（至少在一些商店）。当做一个示范项时，销售经理建议演示一些值得关注的事，就像“啤酒和尿布”被一起卖出。在这个小提示下，分析家能够在数据中发现证据。实际上，该故事的寓意不是关于关联规则的力量，而在于假设测试（hypothesis testing）可能是很有说服力和可操作的。

9.3 一个关联规则有多好

关联规则首先分析包含一种或者多种产品或者服务的交易以及关于交易的基本信息。为了便于分析，称这些产品或者服务为项（item）。表 9-1 展示了一家杂货店包含五种产品的五项交易。

表 9-1 杂货店销售点交易

顾客	项
1	橙汁、汽水
2	牛奶、橙汁、擦窗器
3	橙汁、清洁剂
4	橙汁、清洁剂、汽水
5	擦窗器、汽水

这些交易已经被简化为只包含购买项，如何利用诸如日期和时间信息，以及客户是用现金还是信用卡支付的问题将在本章后面介绍。

每项交易给出了哪些产品会与其他哪些产品一起购买的信息。表 9-2 所示的同现表格（co-occurrence table）显示了这一点，它展示了任何一对产品被一起购买的次数。例如，“汽水”行和“橙汁”列交界处的值是 2，这说明 2 项交易包含汽水和橙汁。这件事情可以很容易地由原来的交易数据验证，其中客户 1 和 4 购买了这两项产品。对角线上的数值（如橙汁行和橙汁列交界处的数值）代表包含该项的交易数目。

表 9-2 产品的同现表格

	橙汁	擦窗器	牛奶	汽水	清洁剂
橙汁	4	1	1	1	2
擦窗器	1	2	1	1	0
牛奶	1	1	1	0	0
汽水	2	1	0	3	3
清洁剂	1	0	0	1	2

这一简单的同现表格已经突出了一些简单的模式：

- 橙汁和汽水更可能比任何其他两项一起被购买；
- 清洁剂从不与擦窗器或牛奶一起被购买；
- 牛奶从不与汽水或清洁剂一起被购买。

这些观察资料都是关联规则的例子，并且可能暗示一个正式的规则，如：“*如果客户购买汽水，那么该客户也购买橙汁。*”眼下，让我们推迟讨论如何自动发现该规则，而是问另

一个问题：这条规则有多好？

在该数据中，五个交易中有两个不仅包括汽水而且包括橙汁。这两个交易*支持*了该规则，对该规则的支持度是 40%。既然两个包含汽水的交易也都包含橙汁，那么规则就有一个高的*置信度*。事实上，三个包含汽水的交易中有两个包含橙汁，于是规则“*如果汽水，则橙汁*”具有 67%的置信度。反过来，规则“*如果橙汁，则汽水*”具有低的置信度。在四个有橙汁的交易中，只有两个有汽水，那么它的置信度只是 50%。更正式地可以说，置信度是支持该规则的交易数目与使规则的条件部分成立的交易数目的比率。另一种表述方式是，置信度是具有全部项的交易数目对只满足“如果”项的交易数目的比率。

另一个问题是该规则与偶然性相比好多少。回答这个问题的一种方式是计算*提升度*（也称为*改善*），它告诉我们在预测结果方面，规则比只是首先假设该结果会好多少。提升度是在应用左边条件之后目标密度与总体中目标密度的比率。表述这一点的另一种方式是，提升度是支持整个规则的记录数与期望数的比率，假定在产品之间没有关系（确切的公式在本章稍后给出）。一个类似的度量，*超额量*，是整个规则支持的记录数减期望值之后的差，因为超额量是用与原始销售相同的单位计量的，有时更容易使用。

图 9-7 提供了一个提升度、置信度和支持度的示例，是由 Blue Martini 公司提供的，这是一家专营零售商工具的公司。他们的软件系统包括了一套包含关联规则的分析工具。这一特定的例子显示，一种特定的外套极可能与一种礼券一起购买，这是一条能够用于改进礼券和外套二者售卖关系的沟通信息。

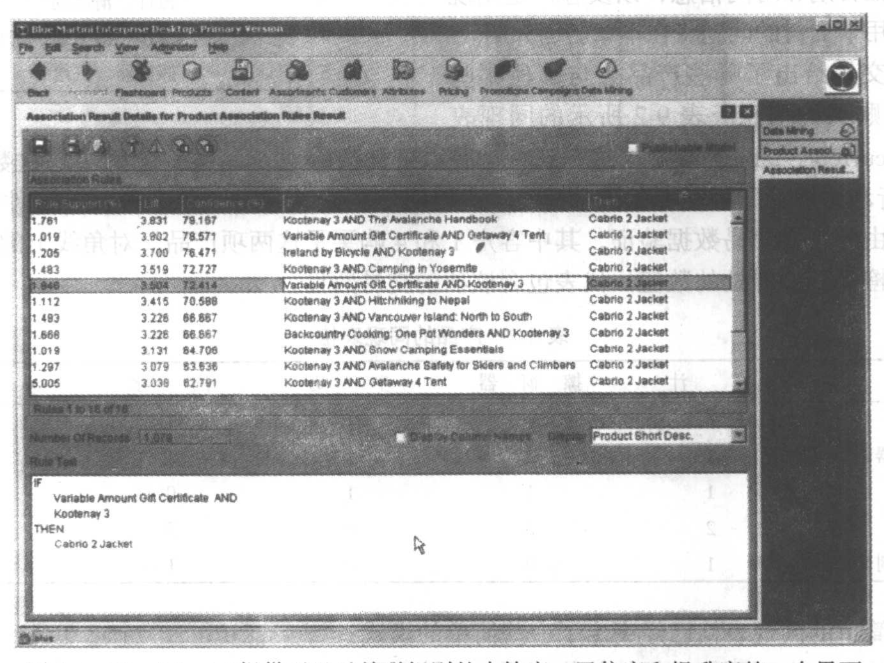
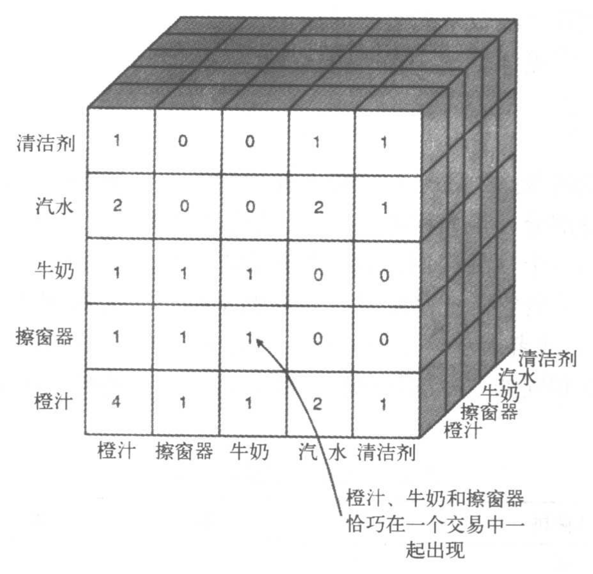


图 9-7 Blue Martini 提供了显示关联规则的支持度、置信度和提升度的一个界面

在同现表格之后的想法是，把组合扩展到任何数目的项，而不只是针对成对的项。对于三种项的组合，可以想象成每个面分成五个不同部分的一个立方体，如图 9-8 所示。即使数据中仅仅有五个项，也已经有 125 个不同的子立方体需要去填充。通过考虑立方体中的对称性，这能降低一点（除以一个为 6 的因子），但由三个项构成的群组，其子立方体数目是不同

项数目的三次幂。一般而言，具有 n 个项的组合数正比于项数目的 n 次幂——一个很快会变得非常巨大的数字，并且产生同现表格需要对每一种组合进行处理。



	橙汁	擦窗器	牛奶	汽水	清洁剂
清洁剂	1	0	0	1	1
汽水	2	0	0	2	1
牛奶	1	1	1	0	0
擦窗器	1	1	1	0	0
橙汁	4	1	1	2	1

图 9-8 一个三维同现表格能够被可视化为一个立方体

9.4 建立关联规则

发现关联规则的基本过程如图 9-9 所示。在创建关联规则时有三点要重点关注：

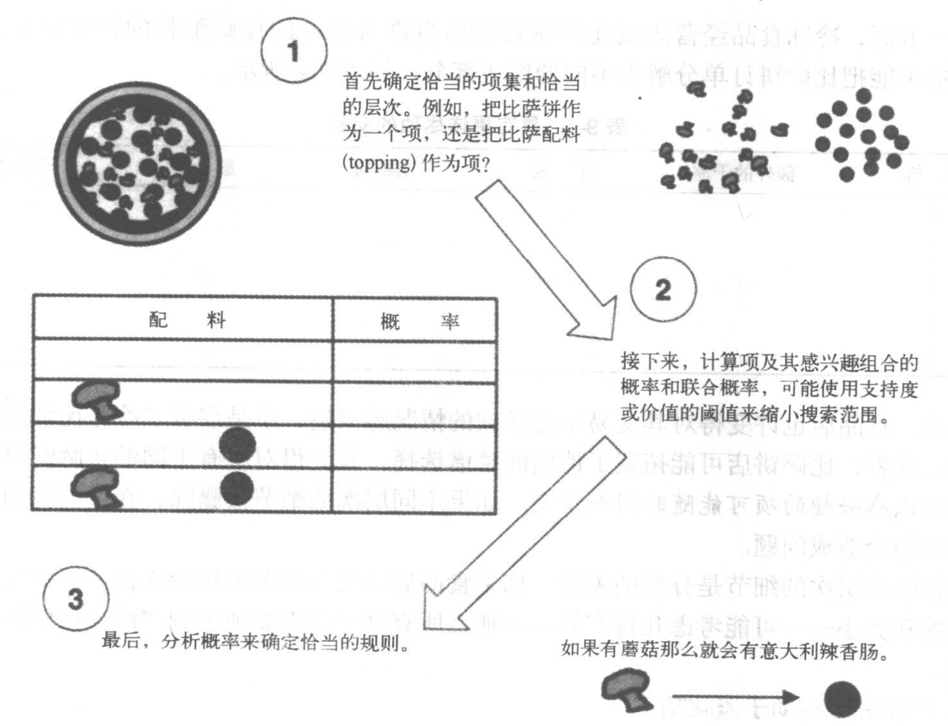


图 9-9 发现关联规则的基本步骤

- 选择恰当的项集。
- 通过判读在同现矩阵中的计数产生规则。
- 克服数千个或数万个项带来的实际局限。

接下来的三个小节将更详尽地深入研究这些要关注的问题。

9.4.1 选择恰当的项集

用于发现关联规则的数据通常是在销售点捕捉的详细交易数据。收集和使用这些数据是应用购物篮分析的关键部分，很大程度上依赖于所选择的用于分析的项。由什么构成特定的项取决于商业需要：在一个食品店内，货架上有数以万计的产品，冷冻的比萨饼可能视为一个分析目标项——而不去管它的比萨配料（额外的干酪、意大利辣香肠或蘑菇）、它的外观（极厚、全麦色或白色）或者大小。于是，购买大的全麦素食比萨饼和购买单份加干酪的意大利辣香肠比萨饼包含相同的"冷冻比萨饼"项。这种交易的物品在汇总层次上看可能如表9-3所示。

表 9-3 有更多汇总项的交易

顾客	比萨饼	牛奶	糖	苹果	咖啡
1	√				
2		√	√		
3	√			√	√
4		√			√
5	√		√	√	√

另一方面，冷冻食品经营店或比萨饼连锁店也许对订购的比萨配料的特定组合很感兴趣。他们可能把比萨饼订单分解为不同的组成部分，如表9-4所示。

表 9-4 具有更详尽项的交易

顾客	额外的干酪	洋葱	胡椒粉	蘑菇	橄榄油
1	√	√			√
2			√		
3	√	√		√	
4		√			√
5	√		√	√	√

后来，食品店也许变得对其交易中更详细的情况感兴趣，于是仅有"冷冻比萨饼"项就不够了。或者，比萨饼店可能拓宽了他们的菜谱选择，并变得对所有不同的比萨配料不太感兴趣，所以感兴趣的项可能随时间而改变。如果不同层次的细节被删除，在试图使用历史数据时这可能会造成问题。

选择恰当层次的细节是分析的关键。如果食品店的交易数据追踪冷冻比萨饼的每一种类型、品牌和大小——可能考虑几打产品——那么所有这些项需要映射到"冷冻比萨饼"项以便分析。

1. 产品分层有助于概化项

在现实世界中，项有产品编码和分成不同层级类别的库存单元代码（stock-keeping unit

code, SKU）（见图 9-10），这种类别称为一个产品分层（product hierarchical）或分类法（taxonomy）。使用哪个产品分层是恰当的？这会带来下列问题：

- 大份炸薯条和小份炸薯条是同一种产品吗？
- 冰淇淋的品牌比口味更具相关性吗？
- 服装的尺寸、格调、款式和设计师哪个更重要？
- 在大型家用电器上的节能选项预示客户的行为吗？

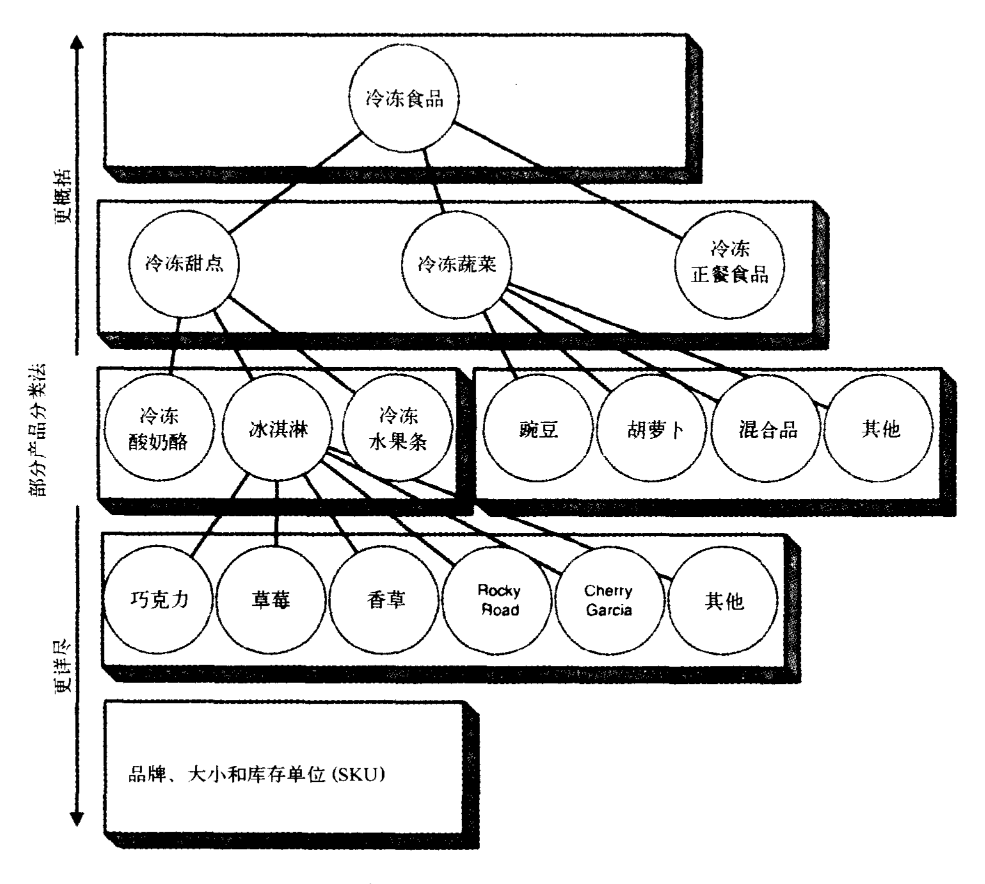


图 9-10 产品分层先概括然后逐步增加细节

当分析中使用的项数增加的时候，要考虑的组合数会非常快速地增长。这要求使用产品分层中较高层次的项，比如使用“冷冻甜点”而不是“冰淇淋”。另一方面，项越具体，结果的可操作性可能越高。例如，知道什么与一种特定品牌的冷冻比萨饼一起销售，有助于处理与生产商的关系。所以，一个折衷的办法是一开始使用概括的项，然后反复生成规则，把目标放到更具体的项上。当分析关注的是更具体的项时，可以只使用包含所需项的交易子集。

规则的复杂性是指它包含的项数。在交易中的项数越多，要产生给定复杂性的规则所花费的时间越长。因此，预期的规则复杂性也决定了项应当多么具体或概括。在某些情况下，客户并不做出大宗购买，例如，在便利店或通过某些目录，客户每次只购买相对很少的几项，因此寻找包含四项或更多项的规则可能只能用于很少的交易，并将会是一项无用的工

作。在另外一些情况下，例如在超市中，平均交易很大，因此更复杂的规则是有用的。

上移产品分层会减少项数，几十个或几百个项可以被减少至单个概化项，常常对应于单个部门或产品系列。例如，一品脱 Ben & Jerry 公司的 Cherry Garcia 这个项可以概化为“冰淇淋”或“冷冻食品”，“橙汁”可以概化为“水果汁”，等等。经常，分层的适当层次会以对应于具有系列产品经理人的一个部门而结束，因此使用分类有发现部门间关系的实际作用。概化项也帮助发现具有足够支持度的规则。分类法中高层次支持的交易数目是低层次支持的交易数目的许多倍。

概化某些项并不意味着所有项需要上移至相同的层次。适当的层次取决于项，取决于其对产生可操作结果的重要性，还取决于它在数据中的频率。例如，在一家百货公司中，高价项（诸如家电）可以停留在层级中的低层次，而低价项（诸如书）可以高一些，这一杂化方法在观察单个产品时也是有用的。因为数据中常常有几千种产品，除了感兴趣的一种或多种产品外，要概化所有其他的东西。

提示：在数据中，当出现在交易中的项数大致相等时，购物篮分析产生最佳结果，这有助于防止规则被最常见的项所支配。产品分层在这里可以有所帮助。把罕见的项滚动到层级中的较高层，于是它们变得频率更高。更常见的项也许根本不必滚动。

2. 虚拟项胜过产品分层

虚拟项的目的是使该分析有能力利用超出产品分层的信息。虚拟项并不出现在原始项的产品分层中，它们跨越产品界限。虚拟项的例子可以是设计师标签，如服饰部门和香水中的 Calvin Klein、食品店中的低脂和无脂产品，以及家用电器上的节能选项。

虚拟项甚至可以包括关于交易的信息本身，诸如该购买用的是现金、信用卡还是支票，该交易出现在星期几或一天中的什么时间。然而，用过多的虚拟项挤满数据并不是个好主意。只有在支持度很好、置信度很高的关联规则中发现可操作的信息，而且有一些使用虚拟项将如何导致这些信息的想法时，才可以包括虚拟项。

但有一个危险，虚拟项能够引起平凡规则。例如，设想有一个虚拟项代表“节食产品”，另一个代表“可乐产品”，那么规则可能如下：

如果“可乐产品”加“节食产品”，那么“节食可乐”

换句话说，＜可乐＞在篮子中出现加上＜节食产品＞在篮子中出现的任何地方，那么＜节食可乐＞也出现。每一个有节食可乐的篮子都满足这个规则。尽管一些篮子也许有普通可乐和其他节食产品，则该规则将有高的提升度，因为它也满足“节食可乐”的定义。当使用虚拟项时，需要检查和复查那些规则以确保没有出现这样的平凡规则。

当右侧不包括关联项时，一个相似的但是更微妙的危险就会出现。于是，规则：

如果“可乐产品”加“节食产品”，那么“脆饼干”

很可能意思是，

如果“节食可乐”，那么“脆饼干”

这种规则的惟一危险是它们能够使正在发生的事情变得不明显。

提示：当应用购物篮分析时，有一个为分析而考虑的项层级分类法是很有用的。通过仔细选择层级的恰当层次，这些概化项在数据中应当出现大约相同的次数，从而改善分析的结果。对于能够深入了解客户行为的具体生活方式的相关选择，诸如无糖项和具体品牌，可以用虚拟项扩充数据。

3. 数据质量

用于购物篮分析的数据通常质量不很高。它是在客户联系点直接收集的，主要用于诸如库存控制的操作目的。数据可能有多重格式、有更正以及不相容的编码类型，等等。对于不同编码值的多种解释可能被深埋在程序代码中，运行于遗留系统中，可能很难提取。同一个连锁店的不同店铺有时有稍微不同的产品分层或不同的打折等处理方式。

这里有一个例子。作者曾经对出现在大型交易数据集中的大约 80 个部门代码感到很好奇，客户使我们确信有 40 个部门，并且对其中每一个都提供了好的描述。通过更仔细的检查我们发现问题之所在。一些店铺用 IBM 收银机，另一些则用 NCR。这两种类型的设备用来表示部门编码的方式不同——因此我们看到在数据中有许多无效的编码。

当使用任何种类的数据进行数据挖掘时，这类问题都是有代表性的。然而，它们增加了购物篮分析的负面影响，因为这种分析很大程度上依赖于未汇总的销售点交易。

4. 匿名与可识别

购物篮分析被证明对大批量市场零售是有用的，诸如超级市场、便利店、药店、快餐连锁店，在那里许多购买者传统上用现金支付。现金交易是匿名的，意思是该店铺不了解具体客户，因为没有信息能识别交易中的客户。关于匿名交易，仅有的信息是交易日期和时间、店铺位置、出纳员、购买的项、兑换的任何优惠券和找零的金额。使用购物篮分析，即使这种有限的数据也能产生重要的和可操作的结果。

渐增的网站交易、忠诚度计划和购物俱乐部的流行，导致了越来越多的可识别交易，使分析师更有可能了解客户信息和客户行为随时间变化的情况。人口统计学和趋势信息可以用于个人和家庭，以进一步扩充客户简档，这一附加信息通过使用虚拟项可以整合到关联规则分析中。

9.4.2 从所有这些数据中生成规则

计算一个给定项的组合在交易数据中出现的次数是适当而有益的，但项的一个组合不是一条规则。有时，仅仅组合自身是有趣的，比如芭比娃娃和糖块的例子。但在另一些情况下，发现一条如下形式的潜在规则更有意义：

如果条件，那么结果。

注意这只是简写。如果该规则说，

如果芭比娃娃，那么糖块。

那么我们把它读作："如果一个客户购买芭比娃娃，那么期望这个客户也会购买糖块。"通常的做法是考虑那些在右侧只有一个项的规则。

1. 计算置信度

构建这样的同现表格，可以提供关于哪种项组合在交易中是最普遍的信息。为便于说明，我们假定最普遍的组合项有三个：A、B 和 C。表 9-5 提供了一个例子，显示各种项和不同组合被购买的概率。

表 9-5 三个项及其组合的概率

组合	概率
A	45.0%
B	42.5%
C	40.0%
A 且 B	25.0%
A 且 C	20.0%
B 且 C	15.0%
A 且 B 且 C	5.0%

只需要考虑具有所有三个项并且在结果中只有一个项的那些规则：

- 如果 A 且 B，那么 C

- 如果 A 且 C，那么 B
- 如果 B 且 C，那么 A

因为这三个规则包含相同的项，它们在数据中具有相同的支持度，5%。它们的置信度水平是多少？置信度是具有规则中所有项的交易数与只有条件中的项的交易数之比。这三个规则的置信度如表 9-6 所示。

表 9-6 规则中的置信度

规　　则	p（条件）	p（条件和结果）	置 信 度
如果 A 且 B 那么 C	25%	5%	0.20
如果 A 且 C 那么 B	20%	5%	0.25
如果 B 且 C 那么 A	15%	5%	0.33

置信度实际表明了什么？如果规则“如果 B 且 C 那么 A”有 0.33 的置信度，则等价于当 B 且 C 在交易中出现时，有 33% 的可能性 A 也在其中出现。换句话说，三次中可能有一次 A 随 B 和 C 出现，另外两次，B 和 C 出现，但 A 没有出现。最可信的规则是最佳规则，因此最佳规则是“如果 B 且 C 那么 A”。

2. 计算提升度

如前所述，提升度是一个关于该规则工作情况有多好的很好度量。它是目标的密度（使用规则的左侧）对总目标的密度的比率。因此公式为：

$$
\begin{aligned}
\text{提升度} &= (p(\text{条件和结果})/p(\text{条件}))/p(\text{结果}) \\
&= p(\text{条件和结果})/(p(\text{条件})\,p(\text{结果}))
\end{aligned}
$$

当提升度大于 1 时，那么得到的规则能更好地预测结果，而不是基于数据中项的频繁程度猜测结果项是否会出现。当提升度小于 1 时，该规则的效果不如按信息猜测好。表 9-7 显示了三个规则的提升度和有最佳提升度的规则。

带有三个项的规则没有一个显示出提升度改善。在该数据中最佳规则实际上只有两个项。在交易中如果购买了“A”，则购买“B”的可能性要比没有购买“A”高出 31%。这种情形和许多情形一样，最佳规则实际上比所考察的其他规则包含更少的项。

表 9-7 四个规则的提升度测量

规　则	支持度	置信度	p（结果）	提升度
如果 A 且 B 那么 C	5%	0.20	40%	0.50
如果 A 且 C 那么 B	5%	0.25	42.5%	0.59
如果 B 且 C 那么 A	5%	0.33	45%	0.74
如果 A 那么 B	25%	0.59	42.5%	1.31

3. 否定规则

当提升度小于 1 时，否定该规则会产生一个好的规则。如果规则：

如果 B 且 C 那么 A

具有 0.33 的置信度，那么规则

如果 B 且 C 那么非 A

就有 0.67 的置信度。由于 A 出现在 45% 的交易中，它就不出现在另外的 55% 的交易中。

应用相同的提升度度量，显示这一新规则的提升度为 1.22（0.67/0.55），得到了一个 1.33 的提升度，比其他任何规则都好。

9.4.3 克服实际局限

生成关联规则是一个多级过程。通常的算法是：

1）对单个项生成同现矩阵。

2）对两个项生成同现矩阵，使用它来寻找有两个项的规则。

3）对三个项生成同现矩阵，使用它来寻找有三个项的规则。

4）依此类推。

例如，在销售橙汁、牛奶、清洁剂、汽水和擦窗器的杂货店，第一步对这些项中的每一项计算计数（count）。在第二步中，创建下列计数：

- 牛奶和清洁剂，牛奶和汽水，牛奶和擦窗器
- 清洁剂和汽水，清洁器和擦窗器
- 汽水和清洁器

这总共是 10 对项。第三步考虑三个项的所有组合，依此类推。当然，其中的每一阶段也许需要一个单独的步骤遍历数据，或者也可以通过同时考虑不同组合的数目，把多个阶段组合到单次遍历中。

尽管在只有五个项时并不明显，但增加组合中的项数需要按指数规律增加计算量，这会导致运行时间呈指数增长——当考虑具有多于三个或四个项的组合时，需要等待更长的时间。解决的办法是修剪（pruning）。修剪是减少每一步考虑的项以及项组合数的一种技术。在每个阶段，该算法抛弃一定数目的不符合某一阈值标准的组合。

最通用的修剪阈值被称为*最小支持度修剪*（minimum support pruning）。支持度指的是在规则支持的数据库中交易的数目。最小支持度修剪需要有一个规则支持最小数目的交易。例如，如果有一百万个交易并且最小支持度是 1%，那么只有被 10 000 个交易支持的规则才是重要的。这是有道理的，因为生成这些规则的目的是继续进行某种行动——诸如和 Mattel（芭比娃娃的生产商）做成一笔交易，制作可食糖块玩偶而且该行动必须影响足够多的交易才是值得的。

最小支持度约束有级联效应。考虑一个有四个项的规则：

如果 A、B 且 C，那么 D。

使用最小支持度修剪，这一规则在数据中必须至少有 10 000 个交易是真的。由此得出结论：

A 必须出现在至少 10 000 个交易中，并且，

B 必须出现在至少 10 000 个交易中，并且，

C 必须出现在至少 10 000 个交易中，并且，

D 必须出现在至少 10 000 个交易中。

换句话说，最小支持度修剪排除了没有出现在足够多交易中的项。阈值标准应用于该算法中的每一步。最小阈值也隐含着：

A 和 B 必须一起出现在至少 10 000 个交易中，并且，

A 和 C 必须一起出现在至少 10 000 个交易中，并且，

A 和 D 必须一起出现在至少 10 000 个交易中，

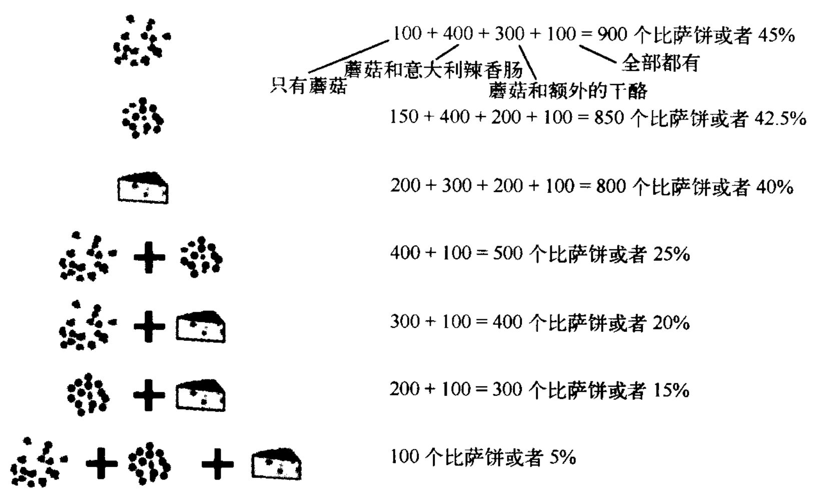
依此类推。

同现表格的每一步计算都能排除不符合该阈值的项组合，减少其大小以及在下一步中要考虑的组合数目。

图 9-11 演示了上述计算如何进行。在这个例子中，选择 10% 的最小支持度水平将把所有具有三个项的组合及其关联规则排除在考虑范围之外。这是修剪在最佳规则上无效的一个例子，因为最佳规则只有两个项。在比萨饼的案例中，这些比萨配料都相当普遍，因此没有被单独修剪。如果在分析之中包括了凤尾鱼——2000 个比萨饼中只有 15 个包含它们——那么 10% 或者甚至 1% 的最小支持度将在第一次遍历中排除凤尾鱼。

一家比萨饼店卖了 2000 个比萨饼，其中：
100 个仅包含蘑菇，150 个是意大利辣香肠，200 个有额外的干酪
400 个是蘑菇加意大利辣香肠，300 个是蘑菇加额外的干酪，200 个是意大利辣香肠加额外的干酪
100 个是蘑菇、意大利辣香肠加额外的干酪
550 个没有额外的比萨配料

我们需要计算项的所有可能组合的概率。



有三个具有全部三个项的规则：

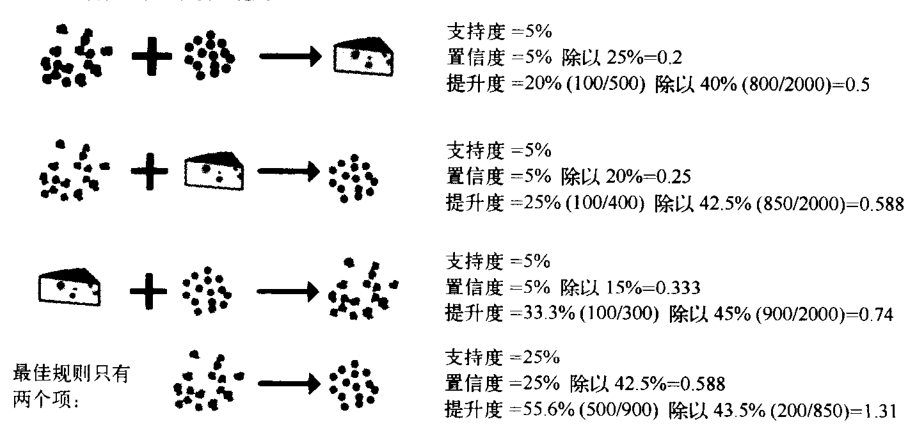


图 9-11 这个例子显示为进行购物篮分析，如何计算比萨饼的销售次数

对最小支持度的最佳选择取决于数据和状况。在算法执行的过程中，也可以变更最小支持度。例如，在不同阶段使用不同水平，能发现常见项的不平常组合（通过逐步减少支持度水平）或者不常见项的相对常见组合（通过增加支持度水平）。

9.4.4 大数据的问题

一个典型的快餐店在其菜单上提供若干项，比如说 100 项。要使用概率生成关联规则，就必须计算每一种项组合的计数。给定大小的组合数倾向于呈指数级增长。一个有三个项的组合可能是小份炸薯条、干酪汉堡包和中份节食可乐。在一个有 100 个项的菜单上，有多少包含三个菜单项的不同组合呢？有161 700个！这一计算是基于二项式定理的。而一个典型的超市的库存中至少有 10 000 个不同的项，更典型的是有 20 000 个或者 30 000 个项。

随着项数在组合中的增加，计算支持度、置信度和提升度迅速变得无法控制。在杂货店中几乎有 5000 万个可能的两个项的组合和超过 1000 亿个三个项的组合。尽管计算机变得越来越强大，要计算这么多组合的计数仍然是非常耗时的。计算 5 个或更多项的计数会昂贵得令人望而却步。使用产品分层可以把项数减少到一个可管理的大小。

交易的数目也非常大。在一年之中，规模适中的超市连锁店将产生几千万或几亿次交易。其中每一笔交易都由一个或更多项组成，常常一次有几十个项。因此，确定项的某个特定组合是否出现在某个特定的交易中可能需要花费一点精力——把所有交易增大 100 万倍。

9.5 扩展思想

关联规则的基本思想能够适用于不同的领域，例如比较不同的店铺，并对规则的定义进行一些强化。这些问题都会在本节中进行讨论。

9.5.1 使用关联规则比较店铺

购物篮分析常常用于对连锁公司的不同分店做出比较。关于五金店中抽水马桶清洁器销售的规则是一个在新店销售和老店销售进行比较的例子。不同的店铺展现出不同的销售模式有许多原因：地域倾向、管理的有效性、不同的广告和在服务范围内各不相同的人口统计学模式。举例来说，在热浪袭击期间人们常去购买空调和风扇，但热浪只影响有限的地域范围。在更小的区域内，服务范围内人口统计学能够有一个大的影响。我们会期望在富人区的店铺与处于穷人区的店铺表现出不同的销售模式。这些是购物篮分析能帮助描述差别的例子，也可以作为使用购物篮分析进行定向数据挖掘的例子。

如何用关联规则做出这种比较？第一步是用*虚拟项*扩充交易，规定交易来自哪个组，例如，老店或新开张的店。尽管该虚拟项不是一个产品或服务，但可以帮助描述这个交易。例如，一家老的五金店销售的物品可能包括下列产品：

- 锤子
- 盒装钉子
- 特细砂纸

提示：*在购物篮数据中加入虚拟交易，则可能发现包括店铺特征和客户特征的规则。*

在扩充规定该交易来自哪里的数据之后，交易看起来像是：

锤子，

盒装钉子，

特细砂纸，

“在一个老的五金店。”

要比较新开张店铺与老店铺的销售情况，过程是：

1）收集新开张店铺在一个规定时期（例如2周）内的数据。增加一个表示该交易来自新开张店的虚拟项扩充数据中的每个交易。

2）从老店中收集大约相同数量的数据。这里可以使用一个跨越所有老店铺的样本，或者可以从位置相当的店铺取得所有数据。用表示该交易来自老店铺的虚拟项扩充这一数据中的交易。

3）应用购物篮分析以发现在每个集合中的关联规则。

4）特别注意包含虚拟项的关联规则。

因为关联规则是非定向数据挖掘，所以规则作为进一步假设测试的起点。为什么一个模式存在于老店铺而另一个存在于新店铺呢？例如，关于抽水马桶清洁器和店铺开张的规则，建议老店在本年内的不同时间更紧密地关注抽水马桶清洁器的销售情况。

使用这一技术，购物篮分析能够用于许多其他类型的对比：

- 促销期与其他时间的销售对比
- 在各种地理区域的销售状况，按照郡县、标准统计都市区域（standard statistical metropolitan area，SSMA）、定向市场营销区域（direct marketing area，DMA）或国家等
- 市区与城郊销售对比
- 销售模式的季节性差别

在每个购物篮中加入虚拟项，使标准关联规则技术能够进行对比。

9.5.2 无关规则

无关规则（dissociation rule）与关联规则类似，只是在条件中用“与非”连接符代替“与”。一个典型的无关规则看上去像：

如果 A 与非 B，那么 C。

无关规则能够通过基本的购物篮分析算法简单改编生成。改编是引入一个新的项集，其中每一项都是初始项的反转项。然后，当且仅当它不包含初始项的时候，修改每个交易使它包括一个反转项。例如，表9-8显示了几个交易的转换。项前面的“-”表示反转项。

表9-8 生成无关规则的交易转换

顾客	项	顾客	加入反转项
1	$\{A, B, C\}$	1	$\{A, B, C\}$
2	$\{A\}$	2	$\{A, \neg B, \neg C\}$
3	$\{A, C\}$	3	$\{A, \neg B, C\}$
4	$\{A\}$	4	$\{A, \neg B, \neg C\}$
5	$\{\}$	5	$\{\neg A, \neg B, \neg C\}$

包含进这些新项有三个负面的影响。第一，用于分析的项总数会加倍。因为计算量按项

数呈指数增长，项的数目加倍严重降低了性能。第二，典型交易的大小增加，因为它现在包括了反转项。第三个问题是，反转项的出现次数趋向于比初始项的出现次数大得多，因此，最小支持度约束倾向于产生所有项都是反转项的规则，例如：

如果非 A 与非 B，那么非 C。

这些规则不像是可操作的。

有时在用于分析的集合中只反转最常出现的项是有用的。当一些初始项的出现频率接近50%时这尤其有价值，这样它们的反转项的频率也接近50%。

9.6 使用关联规则的顺序分析

关联规则发现在同一时间发生的事情——在给定时间购买了哪些项。下一个很自然的问题是关于事件的顺序和它们意味着什么。该类领域的示例有：

- 新的房主在购买家具前购买淋浴帘。
- 购买新的剪草机的顾客很可能在接下来的6周内购买新的橡胶软管。
- 当客户走进银行分支机构并索要账户对账单时，极有可能他或她将关闭其所有的账户。

时间序列数据通常需要一些随时间识别客户的方法。匿名交易不能披露新的房主在买家具之前购买淋浴帘。这要求追踪每个客户，也要知道哪些客户最近买了房子。因为大宗购买常常是用信用卡或借记卡支付的，这很少成为一个问题。对在其他领域的问题，例如调查医疗治理的效果或在银行的客户行为，所有交易通常包含了识别信息。

警告：*为考虑对客户进行时间序列分析，必须找出一些方法识别客户。如果没有方法能追踪单个客户，就不可能分析他们随时间变化的行为。*

针对本节的目的，*时间序列*（time series）是项的有序序列。它不同于仅仅被排序的交易。一般而言，时间序列包含关于顾客的身份识别信息，因为这一信息用于把不同的交易连接到一起组成序列。尽管有许多技术用于分析时间序列，诸如ARIMA（一项统计技术）和神经网络（neural network），但本节只讨论如何把时间序列数据用于购物篮分析。

为了使用时间序列，交易数据必须具有两项额外的特征：

- 一个时间戳或顺序信息来确定交易前后出现的顺序。
- 识别信息，例如账户号码、家庭ID或顾客ID，以识别那些属于同一客户或家庭的不同交易（有时称为一个经济交易单位）。

建立顺序规则与建立关联规则的过程相似：

1）一个客户购买的所有项被当作单个订购处理，每个项保留标记它是何时购买的时间戳。

2）该过程与发现一起出现的项群组的过程相同。

3）为了展开该规则，只有那些左边的项在右边的项之前被购买的规则才予以考虑。

这样得到的结果是一组能够揭示顺序模式的关联规则。

9.7 小结

购物篮数据描述客户购买什么。分析这一数据是复杂的，并且没有一种单一的技术强大到足以提供所有的答案。数据本身通常在三个不同层次上描述购物篮。订单是购买活动的结

果，订单项是购买中的项，客户把订单和时间关联到一起。

关于客户行为的许多重要问题能够通过观察产品销售随时间的变化做出回答。哪些是最佳销售项？哪些项去年卖得挺好而今年不再卖得那么好？存货曲线不要求交易层面的数据，也许它们提供的最重要的信息是市场营销干预的效果——在一个特定的事件之后销售是上升还是下降？

然而，存货曲线对于了解在单个购物篮中项之间的关系是不够的。一项强有力的技术是关联规则，这一技术发现倾向于同时销售的成组产品。有的时候，这个群组对于深入了解事件是足够了；但其他时候，群组被转换为清晰的规则——当特定项出现时，我们期望在篮子中发现某些其他的项。

关联规则有三个度量。支持度反映在交易数据中发现该规则的频繁程度，置信度说明当“如果”部分为真时“那么”部分也为真的频繁程度，而提升度反映该规则预测“那么”部分相对于根本没有规则要好多少。

这样生成的规则可以分成三类：有用的规则阐明可能没有预料到的关系，平凡规则阐明已知（或应该知道）存在的关系，费解的规则没有意义。费解规则常常有很弱的支持度。

购物篮分析和关联规则提供项层次细节的分析方法，其中项之间的关系由它们落入的篮子决定。在下一章中，我们将转向链接分析，它推广了由“关系”链接“项”的思想，利用数学领域中称为图论的内容为背景。

第10章 链接分析

英国航空公司和法国航空公司的国际线路图不仅可以为旅行规划提供帮助，还提供了深入了解各自国家和昔日帝国的历史及政治的相关信息。从纽约启程去蒙巴萨的旅客会在希思罗机场转机，而启程到阿比让的旅客会在高卢的查尔斯机场转机。国际线路图显示出从已知事物的相互联系中能够获得多少信息。

哪些网站与其他哪些网站链接？谁用电话呼叫谁？哪些医师为哪些病人开哪些药？这些关系在数据中都是可见的，并且它们都包含着大多数数据挖掘技术不能直接利用的丰富信息。在联系越来越多的世界中（据说，在这个星球上的任何两个人不存在超过六种程度的分离），理解相互关系和联系是很关键的，链接分析（link analysis）就是定位于这一需求的数据挖掘技术。

链接分析是以称为图论（graph theory）的数学分支为基础的。本章首先回顾图的基本概念，然后展示链接分析如何应用于解决现实问题。链接分析并不适用于所有类型的数据，也不能解决所有类型的问题，但在可以应用的情况中，它常常会产生很富洞察力且可操作的结果。它已经产生很好结果的一些领域是：

- 在万维网上通过分析页面之间的链接识别权威信息源；
- 分析电话呼叫模式，可以识别特定市场群体，诸如在家工作的人们；
- 理解医师转诊介绍模式。转诊介绍是在两个医师之间的某种关系，这又是链接分析非常适用的一个领域。

即使在明确记录链接的情况下，把链接组合成有用的图可能也是数据处理的一个很大挑战：网页之间的链接被编码在页面自身的超文本标记语言（Hypertext Markup Language，HTML）中；电话之间的链接记录在呼叫明细记录中。然而，如果没有相当可观的预处理，这些数据源没有一个对于链接分析是有用的。还有另一些情形，其中的链接是隐含的，数据挖掘要解决的一部分问题就是找到它们。

本章首先简要介绍图论及其解决的一些经典问题，然后转向它在数据挖掘中的应用，诸如搜索引擎分级和呼叫明细记录分析等。

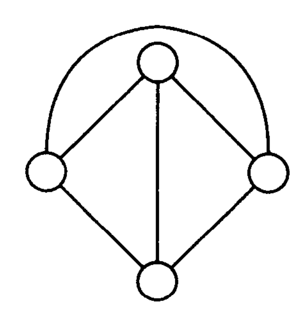
10.1 图论基础

图是专门发展用于表示关系的抽象观念。不论在数学还是在计算机科学中，开发充分利用这些关系的算法已被证明是非常有用的。值得庆幸的是，图十分直观，并有大量的例子阐明如何运用它们。

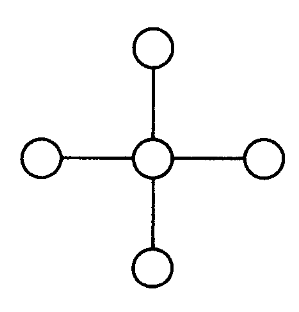
图由两个截然不同的部分组成：

- *结点*（node）（有时称为*顶点*，vertex）是在图中具有某种关系的事物。这些结点具有名称且常常具有额外的有用属性。
- *边*（edge）是通过关系连接的成对结点。一条边可以由它连接的两个结点表示，因此（A，B）或 AB 表示连接 A 和 B 的边。在*加权图*（weighted graph）中边也可能有权重。

图 10-1 给出了两个图。左边的图有四个结点，由六条边连接，具有“在每对结点之间都有一条边”这一特点，这样的图被称为完全连通的（fully connected）。它可以代表在亚特兰大、纽约、辛辛那提和盐湖城之间航线上的日常航班，这四个城市是作为区域交通枢纽；也可以代表全都相互认识的四个人，或者用于刑事调查的四个相互关联的线索。右边的图在中心有一个结点与其他四个结点连接。这可以代表以亚特兰大为中心服务于东南部，连接亚特兰大与伯明翰、格林维尔、夏洛特和萨凡纳等城市的日常航班，或者一家频频被四种信用卡客户光顾的餐馆。图本身捕捉了关于什么与什么相联系的信息，它没有任何标记，可以用于描述许多不同的情形，这就是抽象的力量。



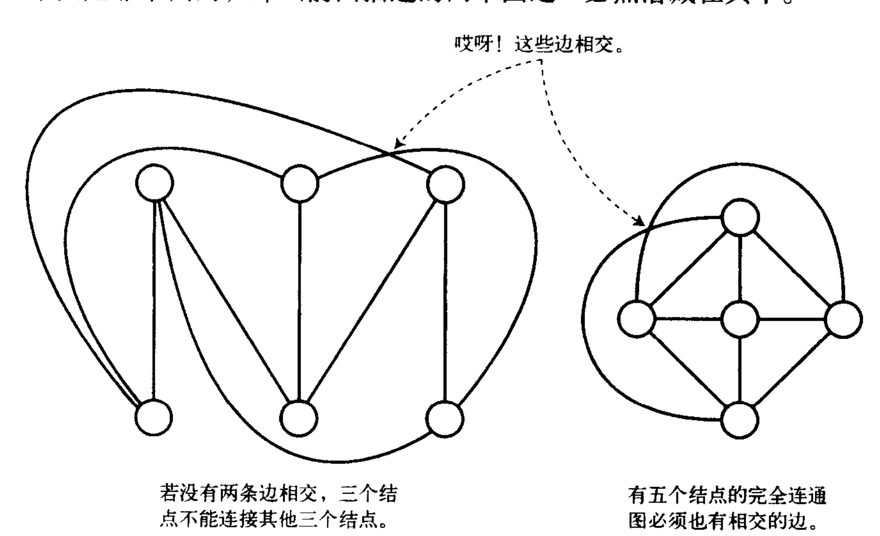
有四个结点和六条边的完全连通图。在完全连通图中，每对结点之间有一条边。



有五个结点和四条边的图。

图 10-1　图的两个例子

关于图有几个术语。因为图对于关系可视化是非常有用的，所以当所有结点与边能够用不相交的边画出时它是完美的。图 10-2 中的图具有这一特性。它们是平面图（planar graph），因为它们能够画在一张纸上（数学家称为平面），且没有任何边相交。图 10-2 显示了两个图，如果没有至少两条边交叉，这两个图是画不出来的。事实上，在图论中有一个定理：如果一个图是非平面的，那么前面描述的两个图之一必然潜藏在其中。



若没有两条边相交，三个结点不能连接其他三个结点。

有五个结点的完全连通图必须也有相交的边。

图 10-2　如果没有一些边互相交叉，不是所有的图都能画得出来

当图中任意两结点之间都存在一条路径时，图被称为连通的。本章后面部分中，除非另有说明，我们假定所有图是连通的。路径，顾名思义，是一个被边连接的结点的有序序列。设想图中每个结点代表一个城市，边是在成对的城市间的航班，在这样的图上，结点是城市，边是航段，而路径是从一个城市到另一个城市的航段路线，诸如从南加利福尼亚的格林维尔到亚特兰大，从亚特兰大到芝加哥，从芝加哥到 Peoria 等。

图 10-3 是一个加权图的例子，其中的边有权重与之相关联。在这一案例中，结点代表顾客购买的产品，边上的权重代表对关联的支持度，即购物篮包含两种产品的百分比。这样的图提供了一种解决购物篮分析问题的方法，它同时也是可视化购物篮数据的有用工具。这个产品关联图是一个无向图的例子。该图显示在这个保健食品店的 22.12% 的购物篮中包含黄胡椒和香蕉两种物品，但这个图本身不能解释是黄胡椒销售激励了香蕉销售，还是香蕉销售激励了黄胡椒，或者是否有另外某些事件驱动所有黄色水果和蔬菜的购买。

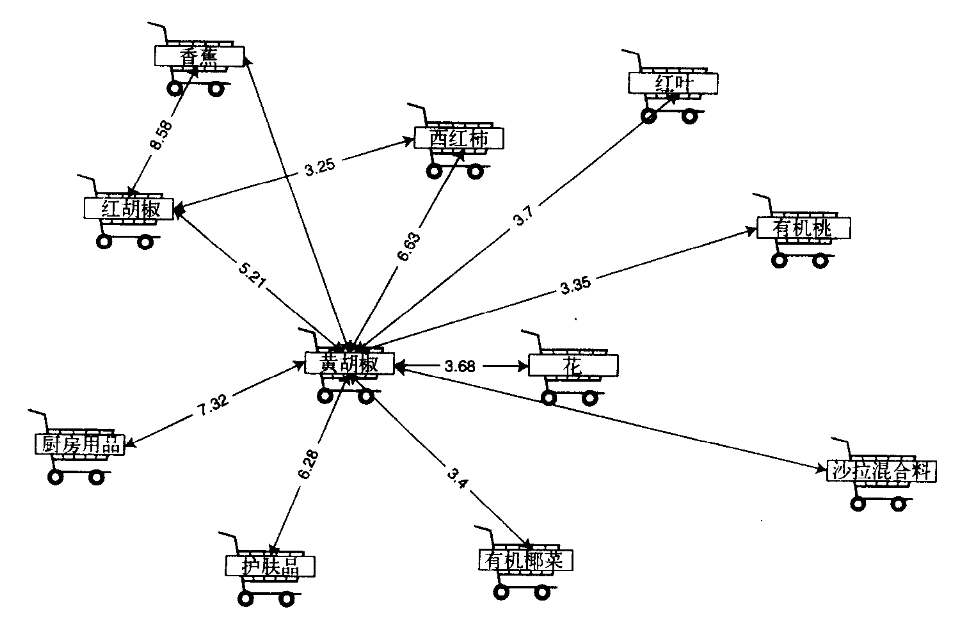


图 10-3　这是一个加权图的例子，其中边的权重是交易的数量，结点的每一端表示购买的物品

在链接分析中一个非常普遍的问题是发现在两个结点之间的最短路径，尽管哪一个最短依赖于指派给边的权重。我们来考虑城市之间的航班图，最短指的是距离吗？还是指最少数目的航段？最短的飞行时间？还是最小的费用？所有这些问题都可以利用图以相同方式回答——惟一不同的是边的权重。

接下来的两节描述了在图论中的两个经典问题，它们展示了图表达问题和解决问题的能力。几乎没有数据挖掘问题恰好与这两个问题相似，但这些问题可以让我们体会到，这些简单图形构筑是如何给出一些重要的解决方案的。给出这些例子的目的，一是通过提供在图论中关键概念的例子使读者熟悉图，二是为讨论链接分析打下坚实的基础。

10.1.1 哥尼斯堡七桥问题

在图论中最早的一个问题起因于一个简单挑战，它是由瑞士数学家莱昂哈德·欧拉在 18 世

纪提出的。如图 10-4 中的简单地图所示，哥尼斯堡有两个岛位于 Pregel 河中，两个岛与城市其余部分之间共通过七座桥相接，在河的任何一边或者在岛上，都有可能到达任何一座桥。图 10-4显示了通过五座桥正好一次穿过该城的一条路径。欧拉提出这样的问题：由城中的任何地方动身，不弄湿身体（译者注：指游过河）或者使用船，有可能一次正好走过所有的七座桥吗？作为一个有历史意义的标记，这个问题已比这个城市的名字存在得还要久。在 18 世纪，哥尼斯堡是座落在立陶宛和波兰之间濒临波罗的海的一个重要的普鲁士城市，现在称为加里宁格勒，是俄罗斯最西部的飞地领土，被立陶宛和白俄罗斯与俄罗斯的其他领土隔开。

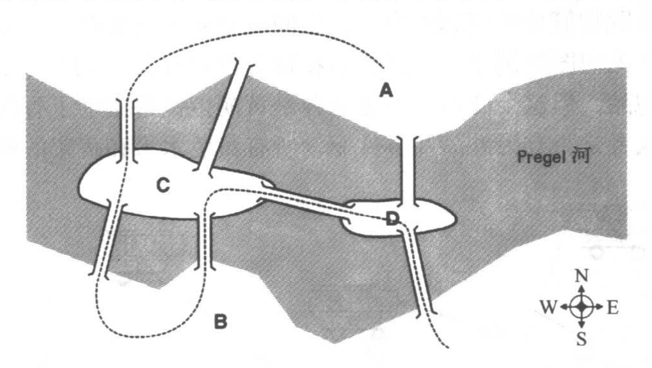


图 10-4　哥尼斯堡的 Pregel 河上有两个岛，由七座桥连接

为了解决这一问题，欧拉发明了图符号表示法。用图 10-5 中所示的具有四个顶点和七条边的简单图来表示哥尼斯堡的地图。一些结点对之间被多条边连接，标志着在它们之间有多于一座的桥。找到一次正好穿过哥尼斯堡所有桥的路线等同于找到在图中一次访问完每一条边的路径。为了向提出和解决这一问题的那位数学家表示敬意，这一路径被称为**欧拉路径**。

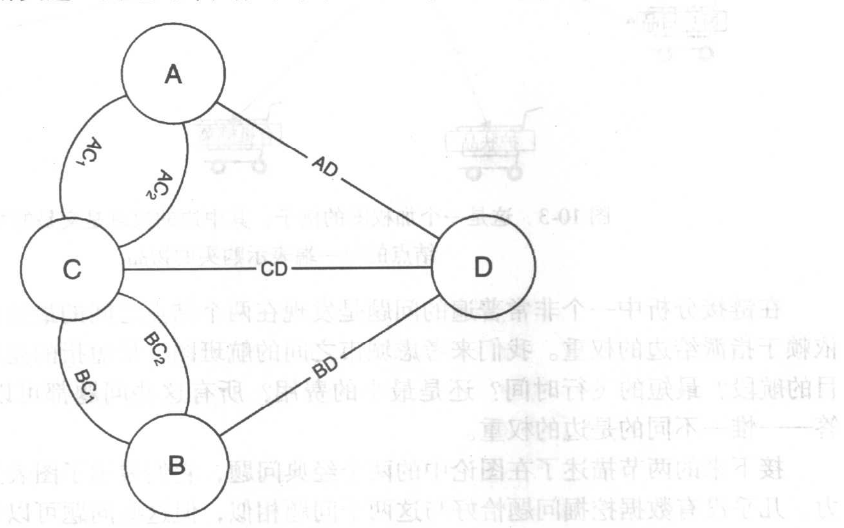


图 10-5　该图表示哥尼斯堡的线路图。边是桥，结点是河岸和岛

为什么度必须是偶数？

依靠简单的观察就可以发现，只有当所有结点上的度是偶数（最多只有两个除外）时，一条欧拉路径才存在。这一观察是关于图中路径的，设想如下一条通过桥的路径：

A→C→B→C→D

使用的边是：

$AC_1 \to BC_1 \to BC_2 \to CD$

图中连接中间结点的边是成对出现的。换句话说，每一条入边都有出边。例如，结点C有四条边访问它，结点B有两条。由于边是成对出现的，每一个中间结点在路径中具有偶数条边。因为欧拉路径包含图中的所有边并访问所有结点，只有当图中所有结点（减掉两个端结点）能充当路径的中间结点时才会存在这样的路径，这是解释那些结点的度是偶数的另一种方式。

欧拉也证明了其反面为真。当图中所有结点（至多两个除外）有偶数个度，那么存在欧拉路径。这一证明有点复杂，但其思想相当简单。要建立一条欧拉路径，可以从任何结点开始（甚至度为奇数的那个），移向任何其他度为偶数的结点。从图上抹掉刚刚经过的边，把它作为在欧拉路径中的第一条边。现在的问题是找到一条起始于图中第二个结点的欧拉路径，当最多有两个结点的度是奇数时，通过保持跟踪结点的度，有可能建立这样一条路径。

欧拉设计了一个基于图中进出每个结点的边的数目的解决方案。这种边的数目被称为结点的度。例如，在表示哥尼斯堡的七座桥的图中，表示海滨的两个结点的度都是3——对应于有三座桥把岛屿连接到对面的陆地这一事实；另外两个代表岛屿的结点具有的度为5和3。欧拉证明，除了最多有两个例外，只有当图中所有结点的度为偶数时欧拉路径存在（见“为什么度必须是偶数”部分）。因此，要走遍哥尼斯堡的七座桥且只能一次穿越一座桥是不可能的，因为有四个度为奇数的结点。

10.1.2 旅行推销员问题

在图论中一个更新的问题是“旅行推销员问题”。在这个问题中，一名推销员需要访问一系列城市中的客户。他打算乘飞机到达其中的一座城市，租一辆车，访问那里的客户，然后驾车到另外的每个城市访问其余的每个客户。他把车留在最后的城市并飞回家。这个推销员能够采用的可能路线有很多。什么路线可以使他旅行的总距离最短而仍然允许他正好一次访问完每个城市？

旅行推销员问题可以很容易地使用图再现，因为图可以很自然地表示被道路连接的城市。在表示这一问题的图中，结点是城市，每条边的权重对应于边连接的两个城市之间的距离。旅行推销员问题因此是在寻求“一次访问图中所有结点的最短路径是什么？”注意这一问题与哥尼斯堡的七座桥有所不同。我们不是对找出恰好一次访问所有结点的路径感兴趣，而是在所有可能的路径中找到最短的一条。注意所有欧拉路径具有完全相同的长度，因为它们包含完全相同的边，寻求最短的欧拉路径没有意义。

对三、四个城市解决旅行推销员问题并不困难。有四个结点的最复杂的图是图中每个结点与余下的每个结点都相连的一个完全连通图，在这个图中，访问每个结点正好一次会有24个不同的路径。要计算路径的数目，从任何结点处开始（有四种可能性），然后走向剩下的三个结点中的任何一个，然后走向其余两个中的任何一个，并最终到达最后的结点（$4*3*2*1=4!=24$）。有 n 个结点的完全连通图具有 $n!$（n 的阶乘）个包含所有结点的截然不同的路径，每条路径都具有稍微不同的边的集合，因此它们的长度通常不同。列出24条可能

的路径不是那么难，对于这一简单的情形，找到最短路径不是特别困难。

找出连接结点的最短路径的问题是爱尔兰数学家威廉·R·哈密尔顿爵士首先提出的。在物理系统中能量最小化研究把他引向某种特定离散系统中能量最小化的研究，他把该离散系统用图表示。为了纪念他，人们把在图中一次访问所有结点的路径称为哈密尔顿路径。

旅行推销员问题难以解决：任何解决方案必须考虑穿越图中所有可能路径，以便确定哪一个最短，而在一个完全连通图中路径的数目增长很快——那是阶乘。对于完全连通图为真的事对于通常的图也为真：访问所有结点的可能路径数目增长是结点数目的指数函数（尽管有一些简单图不是这样）。因此，当城市数目增加时，发现最短路径所需的工作按指数级增长：多增加一个城市（具有关联的道路），就可能导致花费两倍长或更长时间去发现解决方案。

这种可量测性（scalability）的缺乏是如此重要以至于数学家已给它命名 NP——在这里 NP 是指用于解决这个问题的所有已知算法都按指数增加——而不是像多项式那样。这些问题都被认为是困难的，实际上，旅行推销员问题是如此困难以至于它被用于评测并行计算机和奇异算法——诸如用 DNA 或量子物理学诀窍作为计算机的基础，而不是我们更熟悉的由硅制成的计算机芯片。

包含图论在内，对计算机而言有相当好的启发式算法（heuristic algorithm）可以对旅行推销员问题提供合理的解决方案，所给出路径是相对短的路径，虽然并不能保证就是最短的路径。如果你遇到类似的问题，这是个有用的论据。一个通常的算法是贪婪算法（greedy algorithm）：路径起始于图中最短的边，然后用从一端访问新结点的可用的最短边来继续这个路径，这样给出的路径一般是相对短的，尽管不一定是最短的（见图 10-6）。

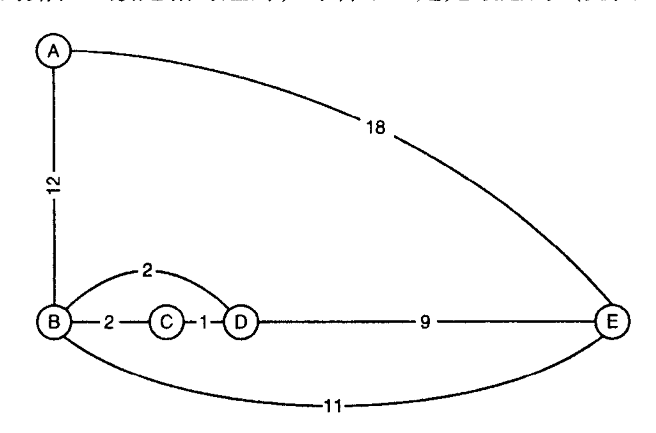


图 10-6　在这个图中，最短路径（ABCDE）的长度为 24，但贪婪算法找到了一条长得多的路径（CDBEA）

提示：通常来说，应该使用一个能得出好的但不是完美的结果的算法，而不是试图分析达到理想解决方案的困难，或者因为不能保证发现最优解决方案而放弃，正如 Voltaire 指出的，“Le mieux est I’ennemi du bien.”（最好是好的敌人。）

10.1.3 有向图

到目前为止讨论的图是无向的。在无向图中，边如同结点之间的高速公路：它们双向通行。在有向图中，边如同单行道，从 A 到 B 的边与从 B 到 A 的边截然不同。从 A 到 B 的有

向边是 A 的出边和 B 的入边。

有向图是表示数据的有效方式：

- 连接一组城市的航段
- 网页之间的超链接
- 电话呼叫模式
- 状态转换图（state transition diagram）

在有向图中，两种类型的结点特别重要。连接*源结点*的所有边是出边。由于没有入边，所以不存在从图中任何其他结点到任一源结点的路径。当结点上所有边是入边时，该结点称为*宿结点*。源结点和宿结点的存在是有向图与无向图之间一个重要的区别。

有向图的一个重要属性是该图是否包含任何起始和终止于同一个顶点的路径。这样的路径被称为环，意思是该路径能够无穷重复自身：ABCABCABC，等等。如果一个有向图包含至少一个环，它被称为*循环的*。例如，航线图中的环可以是单个飞机的路径；在一个呼叫图中，环的成员彼此呼叫——它们或者是会使整个群组得到折扣的“亲友号码组”促销的好候选者，或者是促销会议呼叫服务的好候选者。

10.1.4 检测图中的环

有一个简单算法可以检测有向图是否有环。这一算法首先观察有向图是否不含有宿顶点，且它至少有一条边，然后看是否任何路径能任意延伸。如果没有任何宿顶点，路径的终结结点总是与另一结点相连接，因此该路径能够通过追加那一结点得到延伸。类似地，如果该图没有源结点，那么我们总是能够在路径开头添加一个结点。一旦路径包含比图中结点更多的结点，我们就知道该路径一定至少两次访问了一个结点，把这一结点称为 X。该路径中在第一个 X 和第二个 X 之间的那部分路径是一个环，因此该图是循环的。

现在考虑当图具有一个或更多源结点和一个或更多宿结点的情形。很显然，源结点和宿结点不可能是环的一部分，从图中移除源结点和宿结点，连同它们所有的边，并不会影响该图是否循环。如果这样得到的图不具有宿结点或不具有源结点，那么它包含一个环，就像上面已经讲到的。重复移除宿结点、源结点及其边的过程，直到出现下列情况之一：

- 没有更多的边或没有更多结点留下。在这种情况下，该图没有环；
- 一些边保留下来但没有源结点或宿结点。在这种情况下，该图是循环的。

如果没有环，那么该图被称为一个*非循环图*（acyclic graph）。这些图对于描述事物之间的依赖性或单向关系是有用的。例如，不同产品常常属于能够被非循环图表示的嵌套分层（nested hierarchy）；在第 6 章中描述的决策树是另一个例子。

在一个非循环图中，任意两个结点相互之间具有明确的先后关系。如果在某些包含 A 和 B 二者的路径中，结点 A 先于结点 B，那么在包含 A 和 B 二者的所有路径中 A 都将先于 B（否则可能是一个环）。在这种情形下，我们说 A 是 B 的*前继*（predecessor），B 是 A 的*后继*（successor）。如果没有包含 A 和 B 二者的路径，那么 A 和 B *不相交*。这一严格次序可能是这些结点的重要属性，有时对于数据挖掘目的是有用的。

10.2 链接分析的一个熟悉的应用

本书的多数读者大概都用过 Google 搜索引擎，它的高度普及源于它可以帮助人们找到

几乎关于任何主题的资料的能力，这种功能是通过链接分析完成的。

万维网是一个巨大的有向图。结点是网页，边是页面之间的超链接。称为 Spider 或网络搜索器（web crawler）的特殊程序不断地遍历这些链接，更新网站这一巨大的有向图。这些 Spider 中有些只是简单地索引网页内容以备基于纯文本的搜索引擎使用，而另一些则把网站的总体结构记录为能够用于分析的有向图。

从前，搜索引擎只分析这个图的结点，来自查询的文本与来自网页的文本通过使用类似第8章描述的技术进行比较。Google 的方法（现在已经被其他搜索引擎采用）是不仅利用结点中发现的信息，还利用编码于图的边中的信息。

10.2.1 Kleinberg 算法

一些网站或期刊文章可能比其他一些形式更有趣，即便它们针对的是同一个主题。这个简单的想法很容易领会，但要向计算机解释却很困难，因此，当就许多人写到的一个主题进行搜索时，很难在满足搜索条件的巨大集合中找到最有用或最权威的文档。

Cornell 大学的 Jon Kleinberg 教授提出了解决这个问题的一项广泛采用的技术，他的方法利用了这样的观点：在创建从一个站点到另一个站点的链接中，人类会对被链接到的站点的价值做出判断，到另一个站点的每个链接实际上是对那个站点的推荐。这样累积起来，所有决定链接到相同目标的许多网站设计者的独立判断就是在对那个目标授予权威性。此外，做出链接的站点的可靠性可以由它们链接到的站点的权威性来判断。在决定另一个站点的权威性时，具有许多其他好的推荐的站点给出的推荐能够被赋予更多的权重。

在 Kleinberg 的术语中，链接到许多权威的页面是网络中心（hub）；被许多网络中心链接的页面是权威（authority），这些思想以图10-7说明。这两个概念可以结合起来使用，以分辨“权威”和“仅仅是流行”之间的区别。初看起来，好像发现权威网站的一个好方法是按照无关站点链接到它们的数目对站点进行分级。但这一方法存在的问题是，任何时候，当该主题被一个流行的站点（一个具有许多人站链接的站点）提及，即使是被顺便提及，该网站的权威就会比另一个关于特定主题而较不流行的更权威的站点等级更高。解决的办法是将页面分级，分级依据不是按照指向它们的链接总数，而是按照指向它们的主题相关的网络中心的数目。Google.com 使用了这里描述的基本 Kleinberg 算法的一个改良和增强的版本。

基于链接分析的搜索从一个基于常规文本的搜索开始，这一初始搜索提供一个页面池（常常有两百多个页面），用它开始这个过程。很可能这一搜索返回的文档集并不包括读者将判断为关于该主题的最权威来源的文档，这是因为关于一个主题的最权威来源未必最常使用搜索字符串中的词。Kleinberg 使用了一个用关键字“Harvard”搜索的例子，多数人承认 www.harvard.edu 是关于这一主题的最权威站点之一，但在基于纯内容的分析时，它在一百多万个包含词语“Harvard”的网页中并不突出，因此非常可能的情况是：基于文本的搜索将不会在其结果的前几位返回该大学自己的网站。但很有可能的是：至少返回的一些文档将包含一个到哈佛大学主页的链接；或者如果没有这种链接的话，指向页面池之一的一些页面也将指向 www.harvard.edu。

Kleinberg 算法的一个本质特征是，它不是简单地采纳初始的基于文本搜索返回的页面并试图对它们分级，而是使用它们构造大得多的由根集合中的任何文档指向或被指向的文档池。这个更大的池包含了更多的全局结构——能够被挖掘以确定哪些文档被那些创建池中文

档的人们组成的广泛社团认为是最权威的结构。

10.2.2 细节：查找网络中心和权威

识别权威来源的 Kleinberg 算法有三个阶段：

1）创建根集合。

2）识别候选者。

3）对网络中心和权威分级。

在第一个阶段，使用基于文本的搜索引擎查找包含搜索字符串的页面以形成页面的*根集合*。在第二个阶段，这一根集合扩大为包括被根集合中文档指向或被指向的文档，这一扩展集合包含*候选者*。在第三个阶段，这个过程是迭代的，候选者被依照它们的强度分级为*网络中心*（链接至许多权威文档的文档）和*权威*（有来自许多权威网络中心的链接的页面）。

1. 创建根集合

文档的根集合是使用基于内容的搜索生成的。作为第一个步骤，*无用词*（常用词汇诸如“a”、“an”、“the”等）被从提供的初始搜索字符串中去掉。然后，依据所使用的特定的基于内容的搜索策略，剩余的搜索条件可能经历*词干化*（stemming）。词干化通过移除复数形式和用于动词的各个变化形式、名词格变化等的其他词尾，把词语精简为它们的根形式。然后，搜索网络索引以查找包含搜索字符串中词语的文档。在如何评价匹配的细节上有许多变化，这是为什么在两个基于文本的搜索引擎上执行相同的搜索产生不同结果的一个原因。无论如何，在文档中匹配术语的数目、被匹配术语的稀少程度和提到搜索术语的次数，这些项的组合被用于给索引文档一个确定其关于查询的等级的分数。前 n 个文档用于建立根集合，典型的 n 值是 200。

2. 识别候选者

在第二个阶段，根集合被扩大为创建候选者的集合。候选者集合包括在根集合中*链接至*任何网页的所有网页，加上一个*链接至*根集合中任意页面的子集。如果网络的全局结构作为一个有向图是可用的，那么查找*链接至*特定目标页的页面很简单，同样的任务也能通过使用目标页面的 URL 作为搜索字符串进行基于索引的文本搜索来完成。

仅仅使用链接至根集合中每一个页面的一个页面子集的原因是，防止根集合中的极端流行网站引发难以控制的页面数目的情况发生。还有一个参数 d 可以限制可能被根集合的任何单个成员引入候选者集合的页面的数目。

如果有多于 d 个文档链接至根集合中的一个特定文档，那么 d 个文档的一个随机子集被引入候选者集合，d 的典型值是 50。候选者集合通常将最多包含 1000 到 5000 个文档。

基本算法能够用多种方式改进。例如，一种可能的改进是筛选出来自同一个域内的任何链接，它们中许多可能是纯粹导航性的。另一个改进是允许根集合中的文档从同一站点至多引入 m 个页面。这是为了避免被一个站点的所有页面之间的“合谋”（collusion）所愚弄，例如，网站设计者在每一页面上用“本站设计者为”的链接以广告该站点的情况。

3. 对网络中心和权威分级

最后一个阶段是把候选页面划分为网络中心和权威，并依照它们在那些角色中的强度分级。这一过程也同时有把以下类型的页面分到一组的作用——页面涉及到搜索项的相同内涵但有多种意义——例如，摇滚歌星麦当娜与艺术史中的*圣母和圣婴*（Madonna and Child），

或者美洲虎（Jaguar）汽车与真的大猫一样的美洲虎等。它也能区分出感兴趣主题的权威和一般而言流行的网站的差别：恰当主题的权威页面不仅被许多页面所链接，它们趋向于被同样的页面所链接，正是这样一些网络中心页面把权威结合在一起，使之有别于无关但流行的页面。图 10-7 阐明了网络中心、权威和无关流行页面之间的区别。

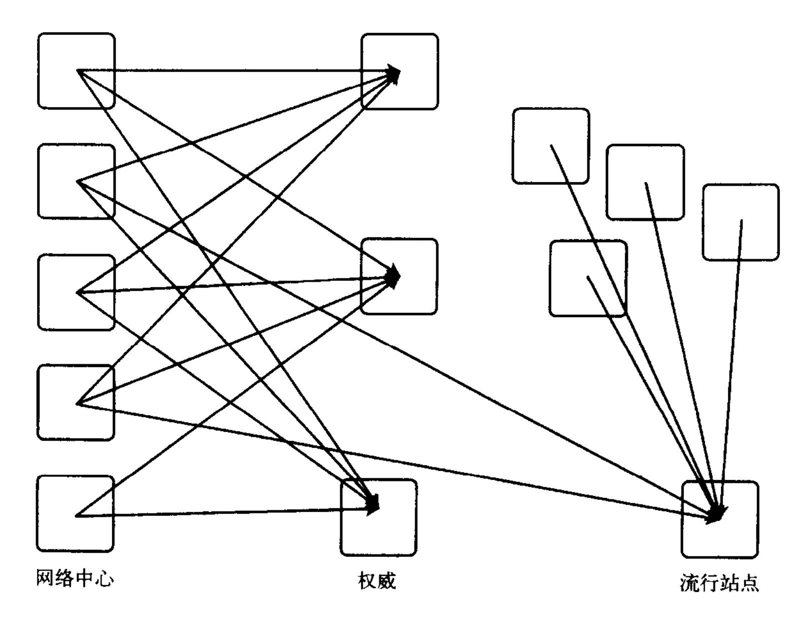


图 10-7 Google 使用链接分析来区分网络中心、权威和流行网页

网络中心和权威具有相互加强关系。一个强大的网络中心链接至许多强大的权威，一个强大的权威也被许多强大的网络中心所链接。算法因此重复进行，首先基于链接到它们的网络中心的强度调整权威的强度，然后基于它们链接到的权威的强度调整网络中心的强度。

对每个页面，有一个值 A 测量其作为权威的强度，值 H 测量其作为网络中心的强度，对所有页面这两个值都被初始化为 1。然后，通过把所有链接到它们的页面的 H 值求和来更新每个页面的 A 值。然后每个页面的 A 值被归一化以便它们的平方和等于 1。然后，H 值以同样的方式被更新。每个页面的 H 值被设定为它链接至页面的 A 值之和，新的 H 值被归一化以便它们的平方和等于 1。这一过程被重复直到 A 和 H 值的一个均衡集合出现。最终，具有最高 H 值的页面是最强大的网络中心，而那些有最强 A 值的页面是最强大的权威。

链接分析的这一应用返回的权威往往是搜索字符串的某个特定含义的强大例子。关于有争议的主题，诸如“同性恋结婚”或“台独”的搜索在正反两方面都产生强大的权威，因为网络的全局结构包括了一些紧密联系的子图，这些子图代表的是具有相似思想的作者所拥护的文档。

10.2.3 实践中的网络中心和权威

关于把链接分析加入到基于文本的搜索的好处，一个最强大案例来自于市场方面。Google，一个由斯坦福大学的 Sergey Brin 和 Lawence Page 开发的搜索引擎，使用一种非常相似于 Kleinberg 的方法，是最早利用链接分析查找网络中心和权威的主要搜索引擎。它很快超越了长期确立的搜索服务，诸如 AltaVista 和雅虎，原因是从质量方面看，它的搜索效

果更好。

2001年4月，当我们研究来自本公司站点www.data-miners.com的网络日志时，注意到了Google返回结果的一些特别之处：当时，行业调查者对网页搜索给了Google和AltaVista大约相等的10%的市场份额，然而Google结果中对我们站点的引用占30%而AltaVista只占3%。这显然是因为Google更能够把我们的站点识别为一个数据挖掘咨询的权威，因为它较少被那些虽然使用“数据挖掘”短语但实际与该主题毫不相干的大量站点所迷惑。

10.3 案例研究：谁在家中使用传真机

图也可以用在来自其他行业的数据中，移动电话、市话和长途电话服务提供商拥有其客户打出和收到的每个电话呼叫的记录，这些数据包含关于其客户行为的大量信息：他们何时发出呼叫，谁呼叫他们，他们是否从其电话套餐中受益，以上仅仅是几个例子。如同这一案例分析所示，链接分析能够被用于分析市话呼叫记录以便识别哪些住宅客户具有在家中拥有传真机的较高概率。

10.3.1 为什么发现传真机是有用的

知道谁拥有传真机有什么用处？一个电话运营商如何按照这一信息采取行动？在这一案例中，提供商已经对在家工作的客户开发了一个服务包，针对营销目的瞄准这样的客户在该公司是革命性的概念。在不久以前还严格管制的市话市场中，本地服务提供者损失了来自在家工作客户的收入，因为这些客户本该支付更高的商业费率而不是较低的住宅费率，因此市话运营商几乎不会去瞄准这样的客户开展市场营销活动，反而可能会拒绝给这样的客户住宅费率——因为它们的行为像小商业企业而惩罚他们。对于这个公司来说，开发和销售在家工作服务包代表一项客户服务的新尝试，但仍存在一个问题：这项新的服务包应当瞄准哪些客户？

有许多方法可以定义客户的目标集合，该公司可以有效地使用地区人口统计学数据(neighborhood demographics)、家庭调查、按邮政编码估计的计算机拥有量以及类似的数据。尽管这一数据改善了市场群体的定义，但离识别具有特定需求的单个客户的要求仍然很远。本书的作者之一曾经所在的一个小组提出，发现住宅传真机使用的能力将改善这一市场营销工作，因为传真机常常（但不总是）用于商业目的。了解谁使用传真机将帮助把在家工作服务包定位到一个很明确的市场群体，与使用基于统计学属性的精确性较差的分段技术所定义的群体相比，这个群体应当具有更好的响应率。

拥有传真机的客户也提供了其他机会。发送和接收传真的客户应当至少有两条线路——如果他们只有一条，就有机会卖给他们第二条线路。为了提供更好的客户服务，在一条有呼叫等待服务的线路上使用传真的客户应当知道如何关掉呼叫等待，以避免传真传送过程中的中断。也有另外的可能性：也许传真机的拥有者更喜欢通过传真而不是邮寄接收他们的每月账单，这样不仅节省邮寄费用也节省打印费用。简而言之，能够识别谁在家中发送或接收传真是有价值的信息，它提供了增加收入、减少成本和增加客户满意度的机会。

10.3.2 用数据画图

用于这一分析的原始数据由某些选定字段组成，它们来自流入账单系统以生成月度账单的呼叫明细数据。每条记录包含80字节的数据，其中含有这样的信息：

- 发起呼叫的 10 位数字电话号码，三位代表区号，三位代表电话局，四位代表线路
- 线路呼叫目标的 10 位数字的电话号码
- 为该呼叫支付账单的线路的 10 位数字电话号码
- 呼叫的日期和时间
- 呼叫持续时间
- 呼叫处于每周的第几天
- 该呼叫是否位于投币式公用电话

在图 10-8 中，数据被缩减至只有三个字段：持续时间、始发号码和终端号码。电话号码是该图的结点，呼叫本身是边，按该呼叫的持续时间加入权重，电话呼叫的一个样本如表 10-1所示。

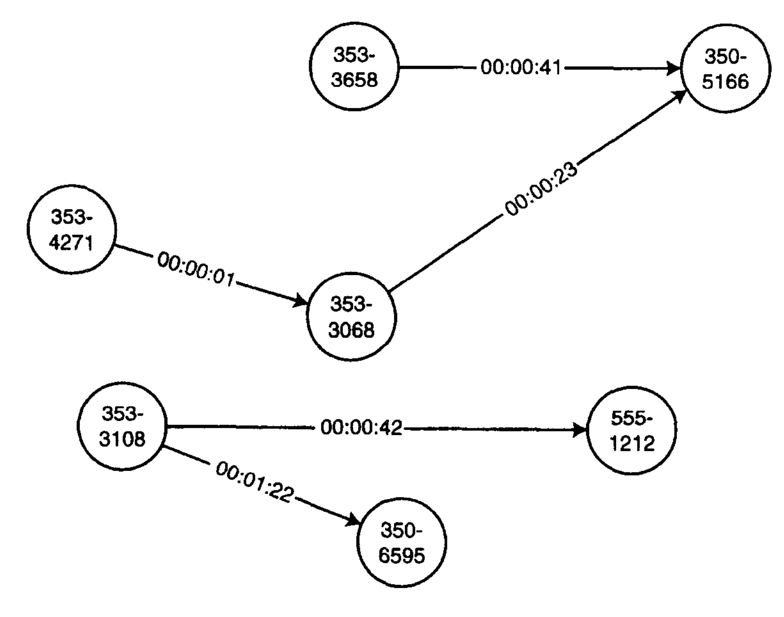


图 10-8 五个呼叫把七个电话号码链接在一起

表 10-1 五个电话呼叫

ID	主叫号码	被叫号码	持续时间
1	353 - 3658	350 - 5166	00:00:41
2	353 - 3068	350 - 5166	00:00:23
3	353 - 4271	353 - 3068	00:00:01
4	353 - 3108	555 - 1212	00:00:42
5	353 - 3108	350 - 6595	00:01:22

10.3.3 方法

查找传真机是基于一个简单的观察：传真机倾向于呼叫其他的传真机。按照从已知号码发出或收到的呼叫为基础，已知传真号码的集合能够被扩展：如果一个未分类的电话号码呼叫已知传真号码并且不很快挂断，那么有证据说明它可能被分类为传真号码。将这一简单的特征

作为指导是很好的，但它过于简化。对住宅客户实际上有几种类型的可预期传真机用法：

- **传真专用线**。传真机的专用线路，该线路只用于传真通信。
- **共享线**。传真机与语音呼叫共享的线路。
- **数据线**。传真机的数据专用线路，或者经由传真或者经由计算机调制解调器。

提示：特征化预期行为是开始任何定向数据挖掘问题的一个好方法。问题理解得越好，结果可能越好。

传真机呼叫其他传真机的假定对于在专用线路上的机器通常是正确的，尽管拨错号码是这一规则的例外。为了把共享线路区分为专用线路或数据线路，我们假定任何呼叫信息台411或555-1212（查号辅助服务）的号码是用于语音通信的，因此是一个语音线路或共享传真线路。例如，样本数据中的第4号呼叫包含一个到555-1212的呼叫，表示该呼叫号码可能是一条共享线路或仅仅是语音线路。当共享线路呼叫其他号码时，没有办法知道该呼叫是语音还是数据，我们不能根据到达或来自呼叫图中这种结点的呼叫来识别传真机。但从另一方面考虑，这些共享线路确实代表了一个销售额外线路的市场营销机会。

用于查找传真机的过程由下列步骤组成：

1）从一组已知传真机开始（从黄页上收集得到），确定一个传真机集合。

2）确定那些向或从这一集合中的任何号码发起或接收呼叫且持续时间大于10秒钟的所有号码，这些号码是候选者。

- 如果该候选号码呼叫过411、555－1212或者一个识别为共享传真号码的号码，那么它被包括到共享语音/传真号码集合中。
- 否则，它被包括到已知传真机集合中。

3）重复步骤1和2直到没有更多号码被识别。

这项工作面临的一个挑战是识别拨错的号码。特别是，到一个传真机的呼入有时可能代表拨错号码，没有给出始发号码的任何信息（实际上，如果它是一个错拨号码那么它多半是一条语音线路）。我们假定这样的呼入错号将持续很短的时间，就像第3号呼叫的情形那样。在一个更大规模的传真机分析中，排除其他例外将是有用的，诸如呼出错号和调制解调器/传真机用途。

该过程始于一个初始传真号码集合。因为这是一个演示项目，几个传真号码是从黄页上根据号码旁边的“传真”注解手工收集的。对一个更大规模的项目，所有传真号码可能从用于生成黄页的数据库中检索得到，这些号码只是传真机电话号码列表的起始、种子。尽管广告其传真号码对于商业很普遍，但这对于家庭中的传真机就不是那么普遍了。

10.3.4 一些结果

电话记录样本由19 674个家庭一个月内的3 011 819个电话呼叫组成。在电话研究领域中，这是一个非常小的抽样数据，但它足以演示链接分析的力量。该分析使用特定的C++代码执行，这种代码存储呼叫明细，并允许我们有效地扩展传真机列表。

查找传真机是*图着色算法*（graph-coloring algorithm）的一个例子。这一类型的算法遍历该图并用不同“颜色”标记结点。在这一案例中，颜色是“传真”、“共享”、“语音”和“未知”而不是红、绿、黄和蓝。最初，除了初始集合中一些标记为“传真”之外，所有结点是“未知”。随着算法的进行，越来越多带有“未知”标记的结点被赋予更多信息的标记。

图 10-9 显示了一个具有 15 个号码和 19 个呼叫的呼叫图，边上的权重是每个电话按秒计的持续时间，对于某个特定号码其实什么都不知道。

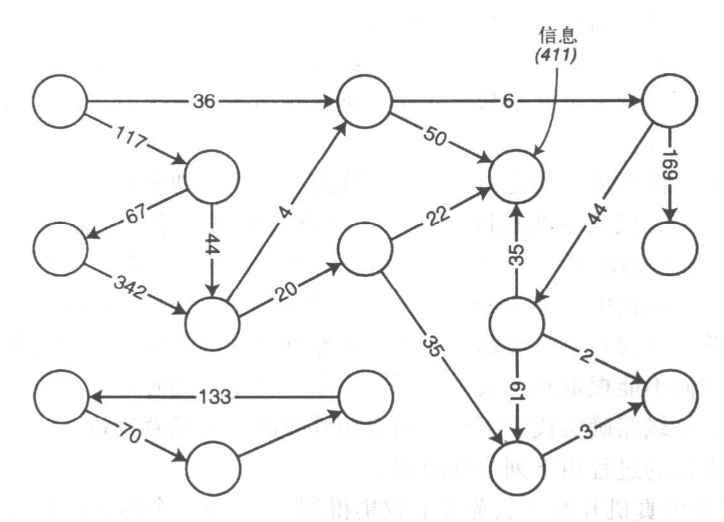


图 10-9 一个有 15 个号码和 19 个呼叫的呼叫图

图 10-10 显示了算法是如何进行的。首先，已知是传真机的号码被标记为“F”，查号辅助号码被标记为“I”，代表持续时间少于 10 秒的呼叫的任何边都被略掉。该算法通过利用一个迭代过程对每一结点指派标记给图着色：

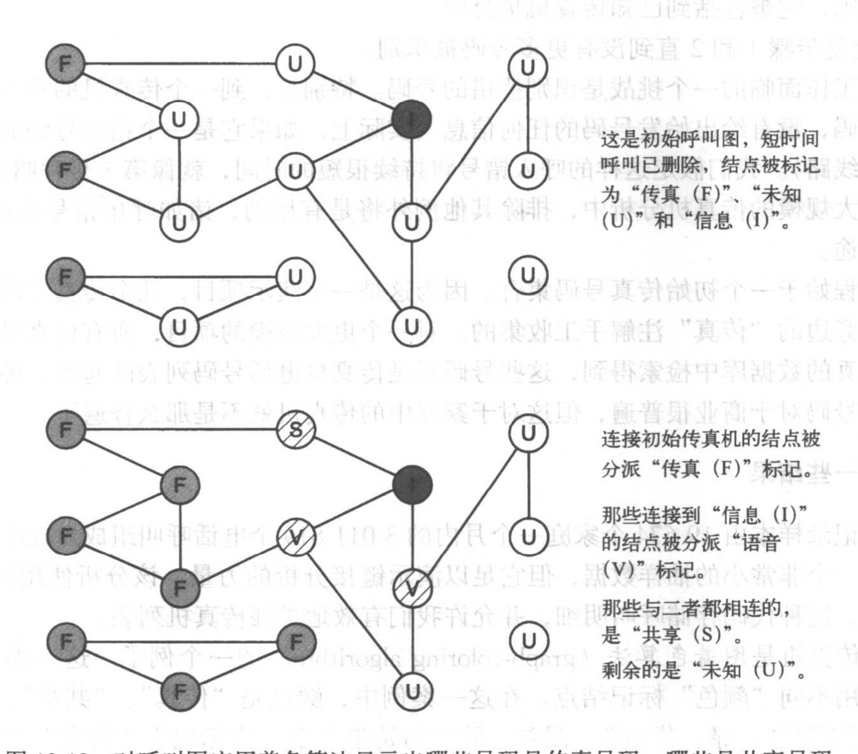


图 10-10 对呼叫图应用着色算法显示出哪些号码是传真号码，哪些是共享号码

- 任何连接到“传真”结点的“语音”结点被标记为“共享”。
- 任何最常连接到“传真”结点的“未知”结点被标记为“传真”。

这一过程持续进行直到连接“传真”结点的所有结点具有“传真”或“共享”标记。

使用 SQL 对图着色

尽管案例分析使用特定的 C++ 代码对图执行着色，但下列操作对于存储在关系数据库中的数据也是适用的。假定有三个表：call _ detail，dedicated _ fax 和 shared _ fax。查找呼叫已知传真号码的查询是：

```
SELECT originating _ number
FROM call _ detail
WHERE terminating _ number IN (SELECT number FROM dedicated _ fax)
  AND duration >= 10
GROUP BY originating _ number;
```

类似的查询能够用于得到由给定传真号码发起的呼叫。然而，这还不能区别专用传真线路和共享传真线路。要做到这一点，我们必须知道是否有任何呼叫是打向信息台的。为了提高效率，最好是把这一列表保存在一个单独的表或视图 voice _ numbers 中，由以下查询确定：

```
SELECT originating _ number
FROM call _ detail
WHERE terminating _ number in ('5551212 ','411')
GROUP BY originating _ number;
```

于是查找专用传真线路的查询是：

```
SELECT originating _ number
FROM call _ detail
WHERE terminating _ number IN (SELECT number FROM dedicated _ fax)
  AND duration >9
  AND originating _ number NOT IN (SELECT number FROM voice _ numbers)
GROUP BY originating _ number;
```

查找共享线路的查询是：

```
SELECT originating _ number
FROM call _ detail
WHERE terminating _ number IN (SELECT number FROM dedicated _ fax)
  AND duration > 2
  AND originating _ number IN (SELECT number FROM voice _ numbers)
GROUP BY originating _ number;
```

这些 SQL 查询是想说明这样的问题：依据关系数据库查找传真机是可能的。它们或许不是针对这一目的的最有效的 SQL 语句，这依赖于数据设计、数据库引擎，以及它运行在什么硬件上。并且，如果数据库中有相当大数量的呼叫号码，用于链接分析的任何 SQL 查询将需要在非常大的表间进行联接。

10.4 案例研究：分段移动电话客户

这一案例分析把链接分析应用于移动电话呼叫，目的是把现有客户分段，以便推销新的服务[⊖]。本节所展示的类似分析方法曾被一家主流移动电话运营商使用，应用从该分析中得到的结果对一种新产品服务进行直接邮寄促销。对于这样的投递，移动电话公司通常测得的响应率为2%到3%，利用这里展示的一些观点，公司把响应率提高到15%以上，这是一个非常显著的改善。

10.4.1 数据

移动电话数据与前面查找传真机的案例分析中看到的呼叫明细数据类似。每一个呼叫有一条记录包含如下字段：

- 主叫号码
- 被叫号码
- 发起该呼叫的地点
- 发起该呼叫的人的账户号码
- 呼叫持续时间
- 时间和日期

尽管该分析不使用账户号码，它在这一数据中仍扮演一个重要的角色，因为没有该数据就不能区分商业账户还是家庭账户。大型商业账户有几千个话机，而大多数家庭账户只有单个话机。

10.4.2 不使用图论的分析

在使用链接分析之前，市场部门曾使用单一度量进行分段：使用分钟数（MOU），即每个月客户移动电话使用的分钟数。MOU是一个有用的量，因为MOU和每个月顾客支付的账单金额直接相关。这一相关是不准确的，因为它没有考虑折扣时段和夜间及周末免费呼叫计划，尽管如此，它仍是一个好的向导。

市场营销部门对潜在客户也有一些外部的人口统计学数据，它们也能用于区分个人客户和商业账户。然而除了MOU之外，他们对客户行为仅有的了解就是支付的总金额和客户是否及时支付账单，他们在表中遗弃了许多信息。

10.4.3 两位客户的对比

图10-11演示了在一个普通月份内两位客户及其呼叫模式。这两位客户具有相似的MOU，然而模式却大相径庭。约翰的呼叫生成一个小的、紧凑的图，而简的呼叫则分解为许多不同的呼叫。如果简非常喜欢无线服务，她的用量将可能增长，并且她甚至可能影响她的许多朋友和同事转到这家无线提供商。

更精密地观察这两个客户会揭示出重要的差异。尽管约翰在车载电话上每个月打到150

⊖ 作者感谢同事Alan Parker、William Crowder和Ravi Basawi对本节所做的贡献。

至200个MOU，但他的移动电话几乎只用在两个用途上：

- 在下班回家的路上，他呼叫妻子让她知道等待时间，有时他们聊三四分钟；
- 每个星期三早上，在早班通勤时间，他在车里进行一个45分钟的电话会议。

惟一有约翰的车载电话号码的人是他的妻子，并且当他驾车时她很少呼叫他。实际上，约翰有另一部携带用于商业目的的移动电话。在驾车时，相对于另一部手提电话，他更喜欢车载电话，尽管他的车载电话服务提供商并不知道这一点。

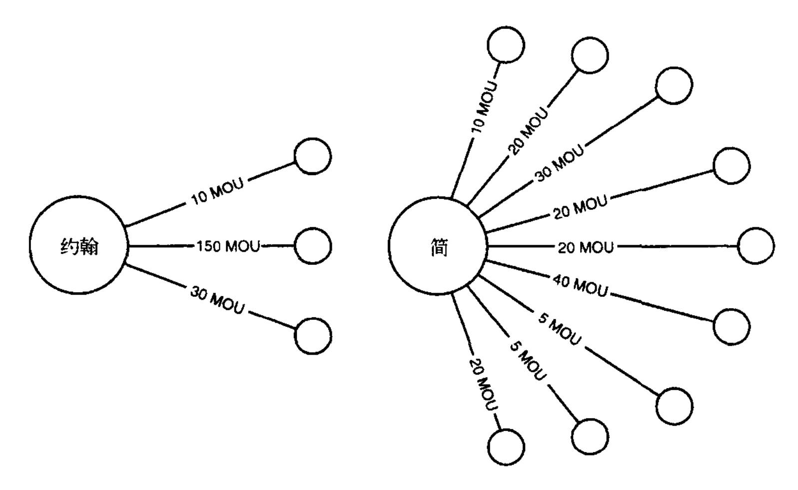


图 10-11 约翰和简具有大约相同的每月使用分钟数（MOU），但他们的行为非常不同

简也在她的移动电话上打大约相同的MOU。她有四个销售人员整天呼叫她，给她汇报情况，当在车里不能找到她时，他们常常在她的移动电话语音信箱中留言。她的呼叫包括对经理人员、潜在顾客和其他同事的呼叫，但呼叫总是非常短——几乎总是一两分钟，因为通常是调度会议。在一家小企业工作，她对个人隐私和呼叫费用很敏感，因此出于习惯，若通话时间长她会使用固定电话。

现在，如果简和约翰都从一个竞争者那里得到一项服务，会出现什么情况呢？谁更可能接受该竞争者的服务（对于该无线通讯公司来说，是*客户流失*）？乍一看，我们可能觉得简对价格更敏感，因此更易转向竞争者的服务。然而，再仔细一想就会发现，如果改变通信公司将需要她改变其电话号码，这对于简是一个大麻烦。（在美国，号码可移植性经历了一个长时间的过程，它于2003年11月终于实现，就在本书这一版出版前不久。这或许会使许多现有的客户流失模型失效。）通过观察打给她的不同人的号码，我们看到简非常依赖于其无线电话号码，她使用如语音信箱及在手机中储存重要号码等功能。如果改变电话号码，她不得不通知很多人，这一点是她保持提供商不变的惯性。约翰不具有这样的惯性，并可能对他的无线提供商不具有忠诚度——只要竞争提供商能对他星期三早上的45分钟呼叫提供不中断的服务就可以了。

简也有许多*影响*。既然她与那么多不同的人通话，他们都会知道她对于移动公司的服务是否满意，她是一个该移动电话公司希望保持满意的客户，但不是一个传统分段方法能够定位的客户。

10.4.4 链接分析的力量

在这一无线电话数据分析中，链接分析扮演两个角色。第一是可视化。可以看出，某些表示呼叫模式的图的能力使得像惯性或影响等这类事物的模式更加显而易见，数据可视化可帮助看到通向更深入问题的模式。对于这个例子，我们选择了早先的分段技术认为相似的两位有利可图的客户，链接分析显示出他们的特定呼叫模式，并暗示客户是如何不同。但从另一方面看，在同一时间观察所有客户的呼叫模式，需要画一个数十万或百万个结点和上亿条边的图。

第二，链接分析能够把通过可视化生成的概念应用于客户的更大集合。例如，减少客户流失的计划可能要避免瞄准具有高惯性的客户，或者确保瞄准具有高影响的客户，这需要遍历该呼叫图，计算出所有客户的惯性或影响，这样得到的特征能够在市场营销工作中扮演重要角色。

不同的市场营销计划可能会建议在呼叫图中寻找其他特征，例如，能够发起电话会议的能力，但谁将是最佳潜在客户？一种想法可能是去寻找全部互相呼叫的客户群组，把它作为图问题表达出来，这一群组是一个完全连通子图。在电话领域中，这些子图被称为“感兴趣共同体”。一个感兴趣共同体可能代表一个对召集电话会议呼叫感兴趣的客户群组。

10.5 小结

链接分析是数学领域中的图论在数据挖掘中的一项应用。作为数据挖掘技术，链接分析具有几种力量：

- 它利用了关系。
- 它对可视化是有用的。
- 它创建能够用于深入挖掘的衍生特征。

一些数据和数据挖掘问题天生包含链接，正像关于电话数据的两个案例分析所示，链接分析对电信是非常有用的——电话呼叫是两个人之间的链接。链接分析可明显用于诸如电话、运输和万维网等链接显而易见的领域，当然，链接分析也适于不具有这样清晰的连接的其他领域，诸如医师咨询模式、零售数据和犯罪的法庭分析。

链接对于可视化某些类型的数据是非常自然的方式。链接的直接可视化对知识发现有巨大帮助。即使自动化模式已经存在，链接的可视化也可以帮助更好地了解正在发生的情况。链接分析提供了观察数据的另一种方法，它不同于关系数据库和联机分析处理（OLAP）工具的形式，链接可能暗示数据中的重要模式，但模式的意义需要由人来解释。

链接分析能够导致新的和有用的数据属性，示例包括对万维网页面计算权威性得分以及对电话用户计算影响范围（sphere of influence）等。

尽管链接分析是强有力的，但它并不适用于所有类型的问题。它不是像神经网络那样的预测工具或分类工具，能够通过输入数据给出答案，许多类型的数据完全不适于链接分析。它的最大作用大概是发现特定的模式（诸如呼出电话的类型），然后将这些模式应用于数据。这些模式可以被转换为数据的新特征，与其他定向数据挖掘技术结合使用。

第 11 章　自动聚类探测

本书中描述的数据挖掘技术是用于寻找有意义的数据模式，但这种模式并不总是立刻就能得到，因为有些时候根本找不到模式；而另一些时候的问题不是缺乏模式，而是模式太多。这些数据可能包含很多复杂的结构以至于最佳数据挖掘技术也不能找出有意义的模式。当挖掘这类数据库以寻找特定问题的答案时，互相对立的解释往往使彼此相互抵消，正像接收无线电信号一样，太多相互竞争的信号叠加到一起就变为噪音。聚类提供了一种获悉复杂数据结构的方法，即将竞争信号的杂音分解成各自的成份。

当人类试图弄清复杂问题的意义时，往往趋向于将问题分解成更小的片段，每一个片段可以更简单地解释。如果要求某个人去描述一片森林中树的颜色，那么在落叶科树和常绿科树之间，在春夏秋冬四季之间，答案可能会大相径庭。人们对林地植物群落已经有足够的了解，可以预知在所有上百种与森林相关的变量中，季节和植物类型是用于按照相似着色规则形成树聚类的最佳因素，它们比植物的年龄和高度等因素更好。

一旦定义了正确的聚类，经常会在各簇之间发现简单的模式，比如“在冬天，落叶树没有叶子，因此树往往是棕色”，或“落叶树叶子的颜色在秋天发生变化，典型的有橙色、红色和黄色”。在许多情况下，非常杂乱的数据集实际上可能由许多表现较好的簇组成，问题是如何发现它们。这就是自动聚类探测技术的用武之地，它可以帮助我们看见整个森林，而不是迷失在树丛中。

本章首先从两个有用的聚类实例开始——其中一个例子来自于天文学，另一个来自于服装设计，然后引入了 K 平均聚类算法，就像在第 8 章中讨论过的最近邻技术，K 平均聚类算法依赖于数据的几何学解释法。将几何学观念应用于 K 平均算法引出了更普通的有关测量相似性、关联性和距离等方面的主题，这类距离测量对于数据的表示方法相当敏感，因此下一个主题讲述的是聚类的数据准备，需要特别注意的是数据的比例缩放和加权问题。K 平均算法不是惟一常用的自动聚类探测算法，本章还对其他几种算法进行了简要讨论，比如高斯混合模型、凝聚聚类和分裂聚类（另外一种聚类技术，也称为自组织映像，在第 7 章中已经学过，是神经网络的一种形式）。本章最后以一个自动聚类探测案例结束，其内容是利用自动聚类探测技术为一家大的日报确定编辑区域。

11.1　搜索单纯岛状片段

在第 1 章，我们把数据挖掘技术分为定向或非定向两大类，自动聚类探测属于非定向知识发现的工具，从技术的角度看情况确实如此，因为自动聚类探测算法本身仅仅是发现存在于数据中的结构，而不考虑任何特定的目标变量。绝大多数数据挖掘任务是从预分类训练集开始，该训练集用于建立模型，对先前未见过的记录给出得分或进行分类。在聚类过程中，没有预分类数据，也没有独立和非独立变量之区别。相反，聚类算法搜寻的是记录的不同分组——即簇（由彼此间相似的记录组成），该算法的目的就是要发现这些相似性。最后，由那些从事分析的人们来确定相似记录是否代表了对商业活动有意义的东西——抑或是某些无法说明的和可能不重要的东西。

然而，从广义角度上看，聚类过程可以看做一个定向活动，因为寻找簇是为了某些商业目的。在营销活动中，针对商业目的而形成的簇通常称为“片段”（segment），而客户分片（segmentation）正是聚类的一项普遍应用。

自动聚类探测是数据挖掘技术中很少被单独使用的技术，因为寻找簇的过程通常并不是数据挖掘的最终目标，一旦找到簇，必须使用其他方法来解释该簇所代表的意义。如果聚类是成功的，结果可能会非常富有戏剧性：聚类探测的一个著名的早期应用导致了目前人们对恒星演变的认识。

11.1.1 星光与星的亮度

20 世纪初期，天文学家试图了解星星的发光度（luminosity，明亮度）和温度之间的关系，他们制作了如图 11-1 所示的散点图，纵坐标以太阳的明亮度倍数来表示发光度，横坐标是以开氏温标表示的表面温度（开氏温标的 0 度称为绝对零度，是理论上可能的最冷温度，用摄氏温标表示的温度值等于开氏温标值加上 273.15）。

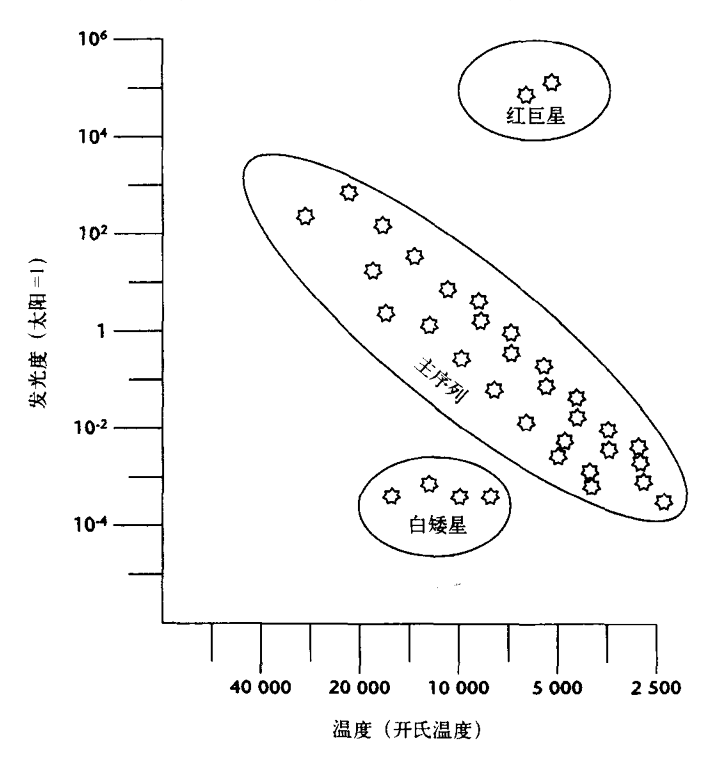


图 11-1 Hertzsprung-Russell 图利用温度和发光度聚类星体

两位天文学家，丹麦的 Enjar Hertzsprung 和美国的 Norris Russell 几乎同时独立地想到这种方法。他们都观察到，在结果的散点图上，星星落到了三个簇中。这一发现引发了他们更进一步的工作，并了解到这三个簇代表了在恒星生存周期中星体所处的完全不同的阶段。在每个簇中发光度和温度的关系是一致的，但各簇间的关系并不相同，因为它们产生热和光的过程根本不同。落在主序列上的 80% 的星星是通过原子核聚变由氢转化为氦产生能量的，这就是所有的星星都要经历的一个活跃的周期。几十亿年后，氢原子会消耗殆尽，依据其质

量的差异，星体或者开始利用氦进行聚变，或者聚变就此停止。在后一种情况中，星星的核心部分就会崩溃，这个过程中会产生大量的热，与此同时，外层气体在远离核心处膨胀，形成一个巨大的火球，最终，外层气体剥离，残留的星核开始冷却，星球变成了一个白色的矮星。

最近在 Google 上用短语"Hertzsprung-Russell Diagram"进行搜索，返回了上千页这类以聚类探测为基础的当前天文学研究相关链接。直到今天，基于 HR 图的聚类仍被用于搜寻褐色矮星（像缺少足够的能量产生聚变物质的星星），以及用于了解主序列时期之前的恒星演变。

11.1.2 适应多维情况

Hertzsprung-Russell 图是一个不错的介绍聚类的例子，因为它只有两个变量，很容易用肉眼发现簇（顺便说一句，这个不错的例子也显示了好的数据可视化的重要性）。甚至在三维空间，从一个立体散点图中用肉眼找出簇也不是很困难。如果所有的问题都只有很少的几个维，就没必要使用自动聚类探测算法了。当维（即独立变量）的数目增加时，发现簇的难度开始增加，我们对于事物相互之间的相近程度的直觉在多维情况下也会迅速瘫痪。

假如一个问题有许多维数，通常暗示需要用几何学方式去分析它。所谓"维"就是用于描述某件事物时需要独立测量的每一个量，换句话说，如果有 N 个变量，就需要设想这样一个空间，其中每个变量的值都代表 N 维空间中沿相应轴的一个距离，N 个变量中的每一个对应于一个值，由所有这些值构成的单条记录可以看做一个矢量，它定义了该空间中的某个特定点。当只有两个维时，很容易画出一个图，HR 图就是这样的一个例子。图 11-2 是另一个实例，绘出的是以一组十几岁青少年的身高和体重为点的图，注意男孩和女孩的簇。

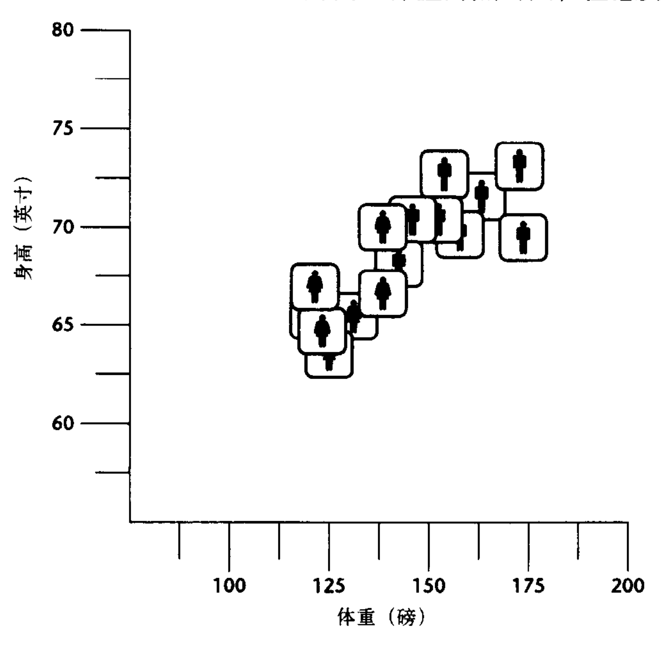


图 11-2 十几岁青少年组的身高和体重

图 11-2 给出了人的体形的一个粗略概念，但如果目的是为了给他们量体裁衣，就需要更多的度量。在 20 世纪 90 年代，美国陆军委托他人研究如何重新设计女兵制服，目的是减

少不同尺码制服的库存数目，但必须保证每个士兵都有合体的制服。

正像任何购买过女性服装的人都注意到的，早已存在纷繁的尺码分类系统（偶数码、奇数码、加大码、年轻的、瘦身的，等等），可用于按照尺码分类服装。但这些系统中没有一个设计是针对美国军队的需求，科内尔大学的研究者 Susan Ashdown 和 Beatrix Paal 从头开始做起，她们基于军队中女性的实际形体设计了一套新的尺码。[⊖]

与传统服装尺码系统不同，Ashdown 和 Paal 所提出的尺码分级系统不是规则地在所有维度上同时变化。相反，她们提出的尺码可以适应各种特定的体型。每一个体型对应于人体测量数据库中一个包含多条记录的簇，某个簇可能由上身长、平均臂长、宽肩、脖子极瘦但腿短、腰细、胸阔的女性构成，而另一些簇则由其他的一组测量数据组成。

数据库中包含了近 3000 个女性的数据，每人有 100 多个度量字段。使用的聚类技术就是下一节要介绍的 K 平均算法（K-means algorithm）。最后，在 100 多个度量数据中只有少数度量用于表征不同的簇，找出这些更少的变量是聚类过程的另一个优点。

11.2 K 平均聚类

K 平均算法是使用最普遍的聚类算法之一，其名称中的“K”指的是算法寻找固定数目的簇，这些簇按照数据点彼此相互接近的程度确定。在这里描述的版本是最初由 J. B. MacQueen 在 1967 年发表的。为了说明方便，我们把这种技术以两维空间图表示。需要牢记的是，在实践中，这种算法通常用于处理多于两个独立变量，这意味着，图中的点不是对应于二元向量（x_1，x_2），而是对应 N 元向量（x_1，x_2，…，x_n）。当然，处理过程本身没有变化。

11.2.1 K 平均算法的三个步骤

第一步，算法随机选择 K 个数据点作为种子（seed），MacQueen 的算法简单地选取从前面数出的 K 条记录，在记录以某种意义排序的情况下，可能需要选择间隔较大的记录，或者随机选择记录。每一颗种子都是仅含有一个元素的基元簇，在这个例子中把簇的数目设置为 3。

第二步，把每一条记录分配给一个最邻近的种子，方法之一是寻找各簇之间的边界，就像图 11-3 中采用的几何方法。两个簇之间的边界就是那些与两个簇等距离的点。回顾一下高中几何课中的一项内容，可以帮助我们更容易地理解这个问题。对于任意两点 A 和 B，所有与这两点等距离的几何点落在一条线上（称为垂直平分线），它垂直于 A 和 B 的连线而且正好平分连线。在图 11-3 中，用虚线连接最初的种子，产生的簇边界用实线表示，它与虚线之间呈一定的角度。利用这些线，可以很容易地看出哪些记录与哪些种子最接近。在三维空间中，这种边界是平面，而在 N 维空间中，就变成 $N-1$ 维超平面。幸运的是，计算机算法可以很容易地处理这种情况。找到簇之间的实际边界有助于从几何角度展示这个过程。在实际工作中，所用的算法经常是测量每一条记录与每一个种子的距离，然后选出最小的距离。

⊖ Susan P. Ashdown 在 1998 年发表了“尺码系统结构的调查：从人体测量数据产生的三个最优多维尺码系统的对比”，发表在《国际服装科学和技术杂志》第 10 卷第 5 期，324-341 页。

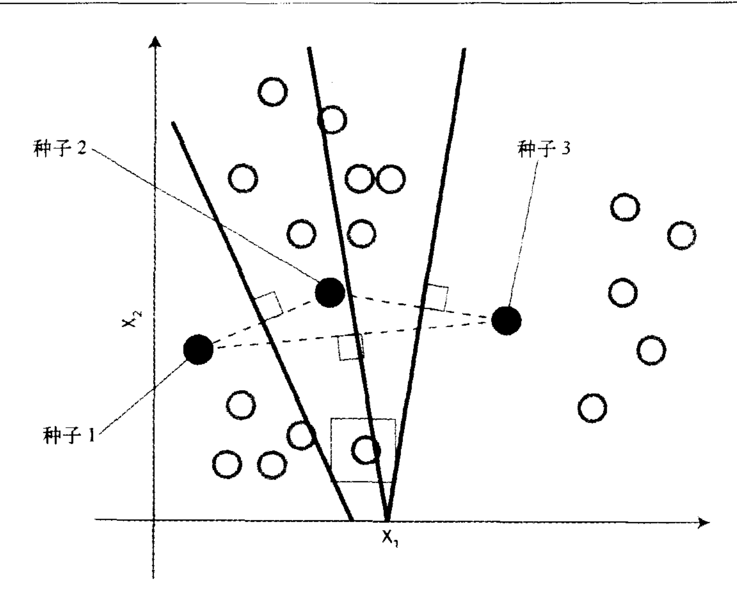


图 11-3 初始种子确定了初始簇的边界

例如，让我们考虑图 11-3 中用一个方框围住一条记录的情况，以最初的种子为基础，这条记录被归入 2 号种子控制的簇，因为它与该种子的距离比其他两个种子都更近。

此时，每一个点都被准确地分配到以初始种子为中心的三个簇之一。第三步就是计算这些簇的形心，这些形心比原来的种子更能代表不同簇的特征。找出形心的方法仅仅是把簇中的所有记录按照维取平均值。

在图 11-4 中，新的形心用“十”字标记出来。箭头所表示的是从初始种子的位置向由这些种子形成的新形心的运动。

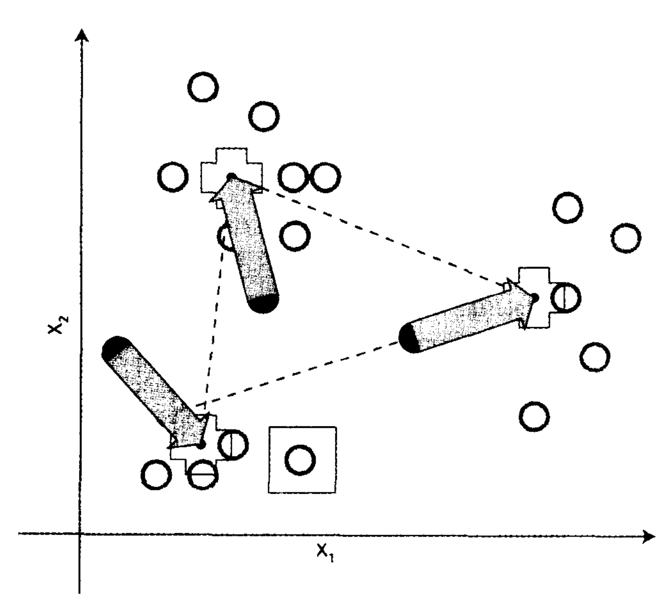


图 11-4 形心是由分配到每个簇中的点计算出来的

这些形心变成了下一次迭代算法的种子，重复第二个步骤，然后每一个点又被归入到离其形心最近的簇。图 11-5 给出了新形成的簇边界，像以前一样，划一条与每个形心等距离的线。注意，被方框围住的那个点原来被分配到簇 2，现在已经被归到了簇 1。这个把点分配到簇然后计算形心的过程一直进行，直到簇的边界不再发生改变为止。实际上 K 平均算法通常在几十次迭代之后才能发现稳定的簇。

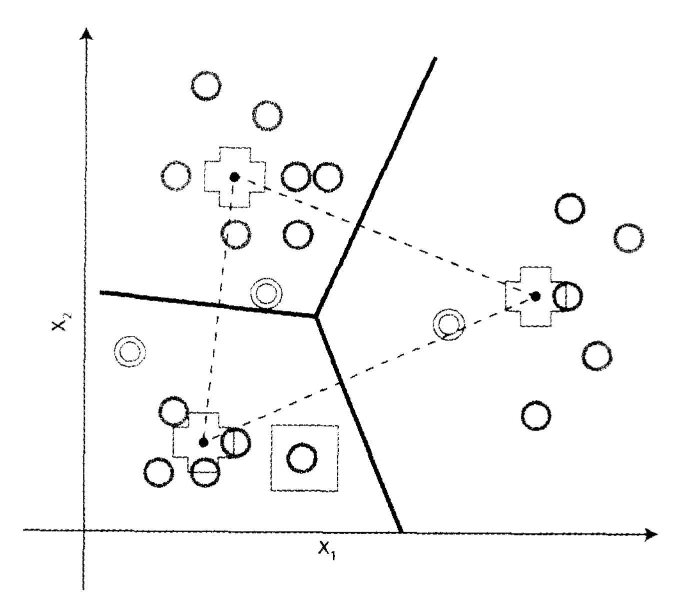


图 11-5 在每一次迭代中，所有簇的分配都经过重新评价

11.2.2 K 的意义

聚类描述了数据中的潜藏结构，但还没有一个对这种结构的恰当描述。比如，某个来自纽约市[⊖]的人可能把整个城市看成“市区”，而另一个来自布鲁克林或昆士的人可能只把“市区”这个词汇用于曼哈顿。而在曼哈顿，它可能只是指第 23 大街以南附近的一片区域，但到了这片区域，“市区”可能只是为曼哈顿岛南端的那些林立的高楼所保留的一个称谓。聚类也存在类似的问题，数据中的结构在不同的层次存在。

对 K 平均和有关算法的描述掩盖了 K 的选择问题，但多数情况下，由于没有一个预先存在的理由去选择一个特定的 K 值，所以对这些算法来说，在进行分析过程中确实存在一个最外层的循环，而这种情况在计算机程序计算过程中反倒不大出现。这种外部循环包括使用一个 K 值进行自动聚类探测，对结果评价，然后用另一个 K 值重新试验或者对数据进行修正。每一次试验后，所得到簇的强度可以通过把一个簇中各记录之间的平均距离与不同簇之间的平均距离进行比较来评价，也可以通过本章中稍后介绍的其他过程进行评价。这些测

⊖ 纽约市（New York City），美国纽约州南部的一个城市，位于哈得逊河口的纽约湾。它是全美国最大的城市和金融、文化、商业、船运和通运中心，最初只包括曼哈顿岛，1898 年重新划定后包括今天的曼哈顿（Manhattan）、布隆克斯（the Bronx）、布鲁克林（Brooklyn）、昆士（Queens）和斯特提岛（Staten Island）五个行政区。——译者注

试可以是自动进行的，但这些簇必须在更加主观的基础上进行评价，以确定它们对于某个应用的实用度。就像图 11-6 显示的那样，K 的不同数值可以导致形成大相径庭但同样有效的簇。图中显示了当 K＝2 和 K＝4 时一副扑克牌的聚类过程，一个聚类是否比另一个更好呢？这取决于聚类将被应用在哪里。

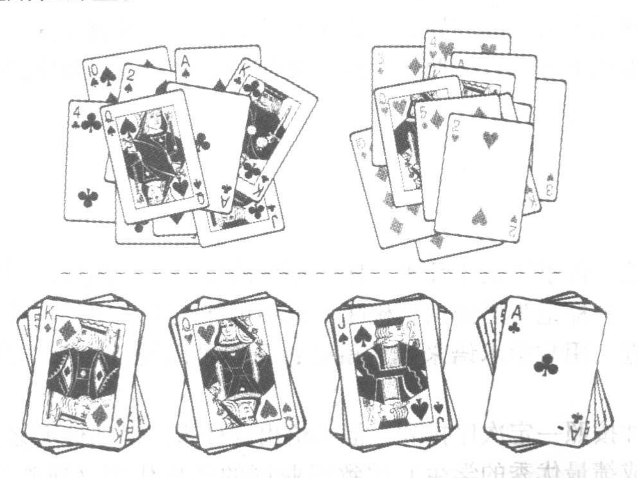


图 11-6　一副扑克牌中，大小为 2 和 4 的簇实例说明，没有惟一正确的聚类

第一次 K 平均聚类（K-means clustering）通常是在给定的数据集上进行，大多数数据点落在一个巨大的中心簇中，而它的外面有许多小的簇。这一切常常是因为在数据中的多数记录表现为“正态”变化，但也有大量外围离群值干扰聚类算法。这一类型的聚类过程或许在识别欺诈行为、制造缺陷等方面的应用上很有价值，而在另外一些应用中，人们可能非常想从数据中过滤出离群值，或者在更多时候，解决方法是修改数据的值。本章稍后有一节讲解为聚类进行数据准备，描述了如何从变量中更容易地找到有意义的聚类方法。

11.3　相似性和距离

一旦数据库的众多记录被定位到空间上的不同点，自动聚类探测确实就是非常简单的问题——一点几何学知识、一些向量均值，这就足够了！然而，问题是，在市场营销活动、销售及客户支持等方面所遇到的那些数据库不是关于空间中的点，而是有关购买、电话呼叫、飞机旅行、汽车登记以及很多其他的事情，它们与聚类图中的点没有一点明显的联系。

对这些类型的记录进行聚类要求有一些自然关联性的概念，就是说，某个给定簇中的记录彼此之间的相似性比与其他簇中记录的相似性更多，因为要把一个直观的概念输送到计算机上是很困难的，所以这个相互关联的含糊概念必须转化为一些描述相似程度的数值型度量，最常用的方法（但绝不是仅有的一种）是将所有的字段转化成数值，这样就可以把记录当作空间中的点来处理，那么，如果从几何学上看两个点是相近的，则在数据库中它们代表的就是相似的两条记录。这种方法存在两个主要问题：

- 许多变量类型，包括所有的分类型变量（categorical variable）和很多数值型变量（比如序列），不适合作为位置矢量中的一部分进行恰当的处理。
- 在几何学上，每一个维的贡献都同等重要，但在数据库中，某个字段上的一点微小的变化可能比在另一个字段上的巨大变化重要得多。

下面的部分介绍几种相似性的替代度量。

11.3.1 相似性度量与变量类型

几何学上的距离对于测量一个良数值型变量的相似性表现良好，“良数值变量”的值表明它在几何模型中数轴上的对应位置，但并不是所有的变量都属于这个范畴。对于这里的目的，可以把变量分为如下四类，按照适应这一几何模型的程度以升序排列：

- 分类变量
- 排序
- 区间
- 真实度量

*分类变量*仅描述一件事情属于几种无序类型中的哪一个，例如，可以把一种冰淇淋标记为开心果味，而另一种标记为奶油山核桃味，但不可能说出其中哪一种更大或者评价出哪一种与黑樱桃味更接近。用数学术语来表达就是：我们可以说 $X \neq Y$，但不能说 $X > Y$ 或者 $X < Y$。

*排序*就是把事件按照一定次序排列，但不给出这件事情比那件大多少。致告别词的毕业生（通常为毕业班成绩最优秀的学生）比致毕业词的学生代表（通常是得第二名）级别更高，但我们不知道高多少。如果 X，Y，Z 按 A，B，C 排序，且我们知道 $X > Y > Z$，但我们不能确定 $X - Y$ 或 $Y - Z$ 的大小。

*区间*度量的是两个观测量之间的距离，如果在旧金山水温是 56 度，而圣何塞水温是 78 度的话，则在海湾的一端比另一端的水温高 22 度。（译者注：这里指的是美国人常用的华氏温度。）

*真实度量*是从一个有意义的 0 点开始测量的区间变量（interval variable），这一特点很重要，因为它意味着两个变量值之比是有意义的。美国使用的华氏温度以及世界其他国家使用的摄氏温度都不具备这个特点，无论上述哪种计量体系，都不可以认为一个 30°的天气会比 15°的天气暖和两倍；同样，一件 12 号的服装不会是 6 号服装的两倍；石膏的硬度也不会是云母的两倍，尽管它们的硬度在硬度表中为 2 和 1。然而确实可以说 50 岁是 25 岁的两倍，10 磅的糖的重量是 5 磅糖的两倍。年龄、重量、长度、客户保有期和体积等都是真实度量的例子。

几何距离度量是有明确定义的，可用于真实度量和区间变量，为了使用分类变量并进行排序，需要把它们转化为区间变量。不幸的是，这些转化可能增加一些伪信息，如果把冰淇淋的口味任意地分配给 1～28 的数字，则看起来 5 号和 6 号口味接近，而 1 号与 28 号味道相差甚远。

上述这些以及其他一些数据转换及准备问题将在第 17 章中详细讨论。

11.3.2 相似性的常规度量

即使没有几百种，至少也有几十种已经公布的技术可以用于测量两条记录的相似性，有些技术是为某些专门用途而开发的，像文本段落的比较；另外一些则是为某些数据类型（如二进制变量或分类变量）特别设计的。对于这里提到的三种情况，前两种适用于区间变量和真实度量，第三种适用于分类变量。

1. 两点之间的几何距离

如果一条记录中的字段是数值型的，则记录表现为 N 维空间中的一个点。对应于两条记录的两点之间的距离常用作它们之间相似性的度量，如果两点间距离相近，对应的记录就是相似的。

有许多方法用于测量两点间的距离，就像下面的“距离度量”部分所描述的，最常用的一种距离是大家熟悉的高中几何中的欧几里得距离，为找出 X 和 Y 之间的距离，可以先找出 X 和 Y 两个点的对应分量（沿每一个数轴的距离）之差，并求出各自平方，平方之和再开方就得到它们间的距离。

距离度量

任何函数，只要能够由两个点产生出一个独立的数值用以描述它们之间的关系，都可以当作一个测量相似性度量的候选者，但要成为一个真正的距离度量，它必须满足以下几个标准：

- 当且仅当 $x=y$ 时，Distance（x，y）$=0$
- 对所有的 x 和所有的 y，Distance（x，y）$\geqslant 0$
- Distance（x，y）$=$Distance（y，x）
- Distance（x，y）$\leqslant$Distance（x，z）$+$Distance（z，y）

这就是几何学中距离度量的正式定义。

真正的距离是寻找簇的一个好度量，但上述某些条件可以有所放宽。其中最重要的条件是第二个和第三个（数学中称为同一性和交换性），即距离为 0 或者一个正值，对任意两点是一个确定的值。如果两条记录的距离为 0，也是允许的，只要它们非常非常相似，因为它们总会落入同一个簇中。

最后一个条件，即三角不等式，从数学角度看也许是非常有趣的。在聚类过程中，它的基本意义是：增加一个新的簇中心不会使两个距离很远的点突然之间看起来靠近了。幸运的是，我们所能设计出的绝大多数度量方法都能满足这个条件。

2. 两个向量间的夹角

有时，考虑两条记录密切关联更有意义，因为其中每条记录的字段都有以某种方式相互联系的相似性。米诺鱼可以与沙丁鱼、鳕鱼及金枪鱼聚成一类；而小猫可以与美洲豹、狮子及老虎聚成一类，尽管用身体各部分长度组成的数据库中，沙丁鱼更接近于小猫而不是鲶鱼。

解决这一问题的方法是对同一个数据使用不同的几何解释，不是把 X 和 Y 作为空间点来测量它们之间的距离，而是把它们当作向量去测量它们之间的*夹角*。在本文的内容中，向量是在坐标系中连接原点及向量值所描述的某一点的一条线段，一个向量既有大小（从原点到点的距离）也有方向，对于这种相似性度量，方向是很关键的。

直接取狮子和家猫的胡须、尾巴、整个身体、牙齿及爪子的长度的数值，把它们各自当作一个独立的点作图，则各点会离得很远，但如果两个物种用身体这些部位的长度求出的比率是相似的，则这些向量几乎线性对应。

向量之间的夹角提供了一种关联程度的度量，它不受两个被比较事物之间数值差异的影响（如图 11-7）。事实上，角的正弦（sine）是一个更好的度量，因为它的取值范围从 0（当向量最接近，即几乎平行的时候）到 1（当它们垂直的时候）。使用正弦函数可以保证处理一个 0°的角同处理一个 180°的角一样，正像应该出现的那样，因为在这种度量中，仅有一

个常数因子不同的任何两个向量都被认为是相似的，即使这个常数因子是负数也是如此。注意，角的余弦（cosine）可以测量相关性，当向量平行（完全相关）时为 1，当垂直时为 0。

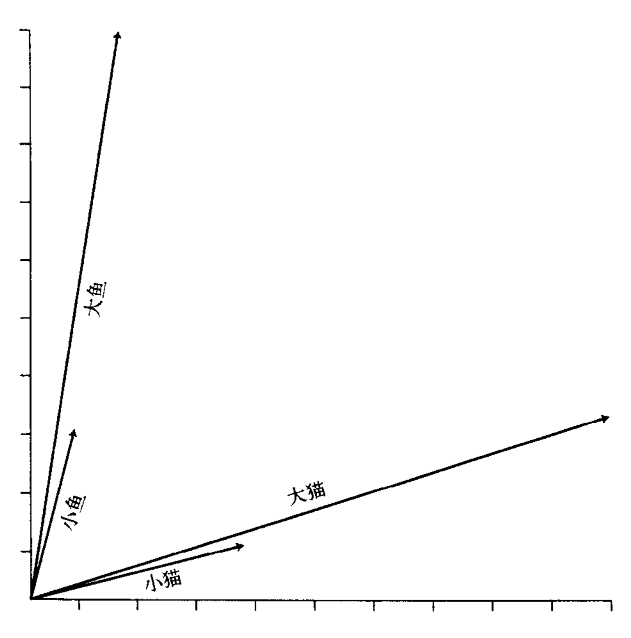


图 11-7 向量间的夹角作为相似性的一个度量

3. 曼哈顿距离

另一个通用的距离度量的名称来源于曼哈顿市中心的街道矩形栅格图，沿着每一条轴线行进可以很容易的对距离求和，这种度量法某些时候成为欧几里得距离的首选，因为沿每个轴线给出的那个距离不被求平方值，不大可能在某一维上产生巨大差异而支配总的距离。

4. 数字的普遍特征

当记录中的字段多数是分类变量时，用几何度量不是最佳选择，一个比较好的度量是基于记录间的重叠程度。与利用几何度量一样，这一概念可以有很多变体，在所有的变体中，把两条记录的字段逐个比较来确定匹配字段数以及不匹配字段数，这样一个最简单的度量就是匹配字段与总字段数的比率。

在这种最简单方式中，这种度量把两个无效或空值字段看做是匹配的，这会产生不正确的结果：使用了缺失数据的每件事情最后都归于同一个簇中。一种简单的改进就是不要包括这些匹配，或者不在匹配计数中包括这种类型；另一种改进是按照总体中某一类型的普遍程度对匹配加权，在“Chevy Nomad”型汽车方面的匹配应当比“Ford F-150”敞蓬小货车占的比重更高。

11.4 聚类过程的数据准备

*比例缩放*和*加权*的概念在聚类过程中都起着重要的作用，二者有些相似，彼此间常常混淆，但二者并不相同。考虑到不同变量是以不同单位或者在不同范围中测量的事实，比例缩放可用于调整变量的值，例如，家庭收入是以数万美元计算的，而孩子的数目是以个位数字表示的。加权方式提供了对变量的一种相对调整，因为其中的一些变量比另一些变量更重要。

11.4.1 利用比例缩放使变量相对一致

在几何学上，所有的维同等重要，在 X、Y 轴上差值为 2、Z 轴方向差值为 1 的两个点之间的距离，与另外两个在 X 方向差值为 1、Y 轴和 Z 轴方向差值为 2 的两个点之间的距离是一样的。在 X、Y、Z 轴上使用什么单位都没关系，只要单位长度一致。

但如果在 X 轴上用“码”作单位，Y 轴用“厘米”，而 Z 用“海里”，那会是什么情况呢？在 Z 轴上 1 的差距相当于 Y 轴上 185 200、X 轴上 2 025 的差距。毫无疑问，必须在距离赋予某种意义之前，把所有值全部转换为一个共同的度量。

不幸的是，在商业数据挖掘中，通常没有一个共同的现成度量可用，因为测量不同的事物使用的是不同的单位，如果变量中包含绘制尺寸、孩子数量、小汽车拥有状况及家庭收入等，它们就不可能全部转换为统一的单位。另一方面，20 英亩的差值与 20 美元的改变之间也是难分辨的，会使人感到费解。一种解决的办法是把所有的变量映射到一个共同的范围（经常是 0 到 1 或 -1 到 1）中，在这样的方法中，至少变化的比率可以进行比较——把一块场地的大小加倍与收入的加倍有同样的效果。比例缩放可以解决这一问题，在这个实例中是把数据重新映射到一个共同的范围。

提示：通过把数值归一化、指数化或标准化，把不同的变量进行缩放，以使它们的值落入一个大致相同的范围，这一点是非常重要的。

以下是三种常用的方法，将变量进行比例缩放后可以全部转换到可比范围。

- 把每一个变量减去最低值后除以范围大小（即最低值与最高值之间的差值），使所有的变量值都映射到 0～1 的范围，这对某些数据挖掘算法是很有用的。
- 把每一个变量值除以变量所有取值的均值，这常被称做“变量指数化”。
- 把每个变量值减去它们的均值，然后除以标准差（standard deviation），这通常称为“标准化（standardization）”或“转换成 z 得分”。一个 z 得分可以告诉你某个值离均值有多少标准差。

归一化一个单独的变量只不过是改变它的范围，一个相近的概念是向量规范化（vector normalization），即一次将所有的变量比例缩放。这也可以用几何学观点解释，把一个单独记录或观测量中的一套数值看做一个向量，归一化过程就是把每一个数值按比例缩放，以便使向量的长度等于 1。将所有的向量转换为单位长度，目的是突出每条记录内在的差异，而不是强调记录之间的差别。作为一个例子，我们来看一条记录的债务和净资产字段，第一条记录包含 $200 000 的债务和 $100 000 的净资产，第二条则有 $10 000 的债务和 $5 000 的净资产，归一化以后，这两条记录看上去是一样的，因为它们的债务和净资产的比率是相同的。

11.4.2 使用权重编码外部信息

比例缩放处理的问题是由于变量数值的大小有差异，使得一个变量的变化看起来比另一个变量的变化更显著。假设考虑这样的情况：两个收入相同的家庭比两个具有同样占地面积的家庭之间的共同点会更多，在聚类过程中如何把这一点考虑进去呢？那就是引入加权，加权的目的是给一个变量是否比另外的变量更（或更不）重要的信息加入编码。

一个好的起点是标准化所有的变量，使每个变量的均值为 0 且方差（或标准差）为 1。

按这种方法，在计算两条记录之间的距离时，所有字段的贡献相同。

我们建议更进一步来考虑问题，自动聚类探测的目的是发现对你有意义的簇，针对要解决的问题，如果人们是否有孩子比他们所带的信用卡数更重要，就没有理由不给孩子字段一个比信用卡更大的权重来突出它在聚类结果中的比重。通过缩放来消除由于单位的不同而引起的偏差以后，再以商业领域的知识为基础利用权重来引入偏差。

某些聚类工具允许使用者对于不同的维增加不同的权重，这就简化了处理过程。即使对于没有这项功能的工具，也可以通过调整比例缩放值来加入权重。方法是，首先把该值缩放到一个常见的范围以消除范围带来的影响，然后根据商业环境，把得到的结果乘以一个权重来引入偏差。

当然，如果你想评价不同加权策略带来的影响，就需要在聚类过程中增加另一个外层循环。

11.5 聚类探测的其他途径

基本的 K 平均算法有许多变体，包含自动聚类探测的许多商业软件工具吸收了这些变体中的某一些算法，它们的差异之处在于挑选初始种子的不同方法以及概率密度使用，而不是通过距离把簇中的记录联系起来。上述方法的最后一种变体值得进一步讨论，此外，聚类过程可以有几种不同的方法，包括凝聚聚类、分裂聚类及自组织映像等。

11.5.1 高斯混合模型

正如人们指出的，K 平均算法有一些缺点：

- 它对于有重叠的簇表现不大好；
- 簇很容易被离群值牵引而偏离中心；
- 每条记录只有一种情况，就是属于或者不属于一个已知簇。

*高斯混合模型*是 K 平均算法的概率论变体，这个名字来源于高斯分布，这是一个常用于高维问题的概率分布。高斯分布推广了多于一个变量的正态分布，像以前一样，该算法也是从挑选 K 个种子开始。但是，这一次的种子是高斯分布的均值，该算法过程是通过被称作估计步骤和最大化步骤的两步反复进行的。

估计步骤为每个数据点的每一个高斯模型计算出*响应度*（见图 11-8），在每个高斯模型中，那些靠近均值的点有高响应度值，而远离均值的点有低响应度值，所以响应度在下一个步骤中被当作权重使用。

在最大化步骤中，把这些新计算出的响应度考虑进来，为每个簇计算出一个新的形心，某个高斯模型的形心可以通过对该高斯模型的所有点的*响应度加权后*取平均来求出，如图 11-9 所示。

重复这些步骤直到高斯曲线不再移动，当然，像位置的移动一样，高斯曲线本身的形状也可以发生变化，然而，每条高斯曲线是受约束的，因此如果靠近均值的点有很高的响应度，那么其响应度必然在其他地方会有明显的谷。如果高斯曲线覆盖了大的数值范畴，那么它附近的点会有较小的响应度，由于该分布必须始终保持积分为 1，所以当高斯曲线范围变大时其强度会变弱。

称它为“混合模型”的原因是在每一个数据点上的概率是几个分布混合后的总和。在这

个过程的最后，每个点被绑定到一个具有高或低的概率的不同簇上，这有时被称为软聚类(soft clustering)，因为其中的点不惟一对应于单个簇。

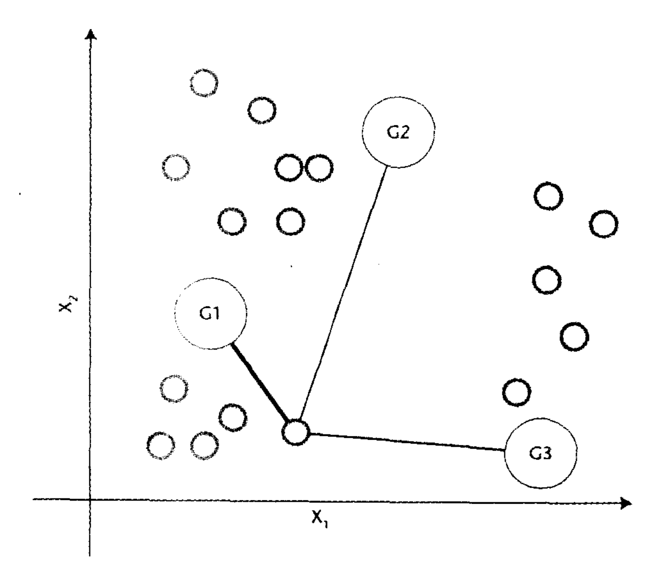


图 11-8 在估计步骤中，对每个点都为高斯模型分配了一些响应度，粗线代表高响应度

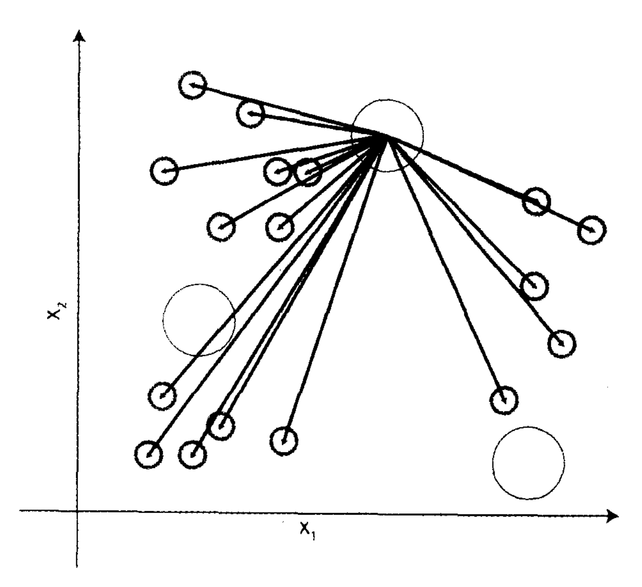


图 11-9 每个高斯均值被移到所有数据点的形心，每个点的响应度被加权，粗箭头表示高权重

这种方法的一种结果是一些点可能在多于一个簇中有高的概率，另一些点在每一个簇中可能都有很低的概率，每个点可能被分配到概率最高的簇，这样就把这些软聚类转化为硬聚类（hard clustering）。

11.5.2 凝聚聚类

用 K 平均算法进行聚类首先是从一个固定数目的簇开始，将所有的记录分配到这些簇

中。另一种分类法是通过凝聚来聚类，这种方法是从每个数据点构成它自己的簇开始，逐渐将它们合并成越来越大的簇，直到所有点凝聚到一起成为一个大的簇。在这一过程的开始阶段，簇很小且很纯——每个簇的成员很少且密切相关，而在这一过程结束时，簇很大且界限不分明。整个过程都被保存下来，以便可能从中选出针对给定应用能得出最好结果的聚类层次。

1. 凝聚聚类算法

第一步是创建一个*相似性矩阵*（similarity matrix），该相似性矩阵是由簇两两之间的距离或相似程度形成的一个表格。最初，相似性矩阵包含了单独的记录对两两之间的距离。就像前面已经讨论的，有很多关于记录之间的相似性度量，包括欧几里得距离、向量间的夹角及分类字段之间匹配与不匹配比率等。距离度量的选择引出的这些问题与先前关于K平均方法讨论的相同。

这可能像 N 个数据点有 N 个初始簇的情况，需要有 N^2 个测量计算来创建距离表。如果相似性度量是一个真实距离度量，只需做一半工作就可以了，因为所有真实距离度量遵循下列规则：Distance（X，Y）=Distance（Y，X）。在数学意义上，相似性矩阵是倒三角形的。下一个步骤是寻找相似性矩阵中的最小值，这可以用于识别两个最相似簇，将这两个簇合并为一个新的簇，把描述父簇的两行替换成描述合并后的簇与剩余簇之间距离的一个新行，以此更新这个矩阵，现在相似性矩阵中有（$N-1$）个簇和（$N-1$）行。

重复（$N-1$）次上述合并步骤，则所有记录都属于同一个大簇。每一次迭代都记住哪些簇被合并，这些簇之间的距离是多少，这些信息常用于确定到底使用哪个层次的簇。

2. 簇间距离

关于如何测量簇间距离还需要稍微多讲一点。在整个合并步骤的最初阶段，簇只包含单条记录。因此，簇间距离等同于记录之间的距离，这可能是在前面已经详细介绍的一个主题。在该循环的第二和后续阶段中，需要利用新的多记录簇与其他所有簇的距离来更新相似性矩阵。我们到底该如何测量这个距离呢？

像通常一样，需要选择一个方法，常用的三种为：

- 单一链接
- 完全链接
- 形心距离

在*单一链接方法*（single linkage method）中，两个簇间的距离是由两个最接近的成员的距离给出的，这种方法产生的簇有这样的特性：与处于簇外面的任何一个点相比，簇中的每个成员都与簇内部至少一个成员在距离上更接近。

另一种方法为*完全链接方法*（complete linkage method），两个簇之间的距离是由最远成员之间的距离给出的，这种方法所产生的簇的特性是：所有成员彼此位于一个已知的最大距离内。

第三种方法为形心距离，两个簇之间的距离是通过每一个簇的形心之间的距离来测量的，一个簇的形心是它成员的平均。图11-10给出了这三种方法的示意图。

3. 簇和树

凝聚算法创建分层的簇，在每一层上，簇由下一层的两个簇合并在一起形成，可视化这些簇的一个好方法是利用树，当然，这样的一棵树看起来可能像第6章中讨论的决策树，但有很大的差别，一个最大的差别是，簇树的结点中并不嵌入规则来描述为什么会产生聚类；

这些结点仅表示一种事实，即两个子结点是所有可能的簇对中距离最小的。另一个差别是，创建决策树的目的是最大化给定目标变量的叶纯度，而簇树却没有目标，只是表示每个簇内的自相似。本章稍后的部分将讨论分裂聚类方法，它与凝聚聚类相似，只是凝聚法是由叶子到根部生成聚类，而分裂法是由根部到叶子生成聚类。

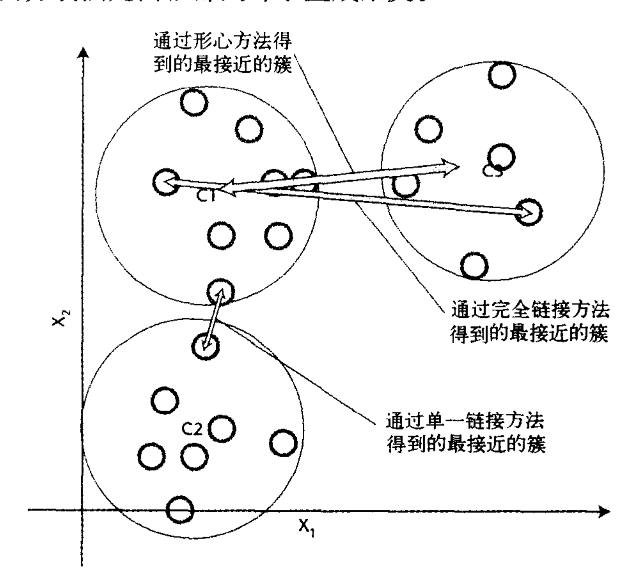


图 11-10 测量簇之间距离的三种方法

4. 通过年龄对人进行聚类：凝聚聚类示例

这个凝聚聚类的说明使用了在一维空间中两个簇之间的距离度量的单一链接实例，这些选择使它可能在整个迭代过程中遵循该算法，而不必担心使用平方和平方根对距离进行计算。

这些数据由许多以家庭为单位收集来的人的年龄构成，目的是使用他们的年龄找出这些参与者的簇，两个人之间距离的度量就简单地用他们的年龄差。人群的两个簇之间的距离标度采用的也是一个年龄差，即较年轻的簇中最年长成员的年龄与较年长的簇中最年轻的成员之间的年龄之差（单一链接测量的一维空间版本）。

因为该距离很容易测量，这个例子就无需用相似性矩阵，其过程是把参与者按照年龄分类，然后开始聚类过程，首先把相差 1 岁的簇合并，接着是相差 2 岁的，依此类推，直到仅剩下一个大簇。

图 11-11 显示了 6 次迭代之后保留下来的三个簇，这是看上去十分有用的聚类层次。算法似乎把人口聚类到三代中：孩子、父母及祖父母。

11.5.3 分裂聚类

我们已经注意到，在凝聚聚类技术形成的树与决策树算法形成的树之间有某些相似性，虽然凝聚方法的工作是从叶到根，而决策树算法的工作是从根到叶，但它们都产生一个相似的分层结构。这种分层结构反映了两种方法的另一个相似之处：先前过程中做出的决定从不被回访，这意味着，如果早期的分裂或者凝聚会破坏该结构的话，某些相当简单的簇可能会探测不到。

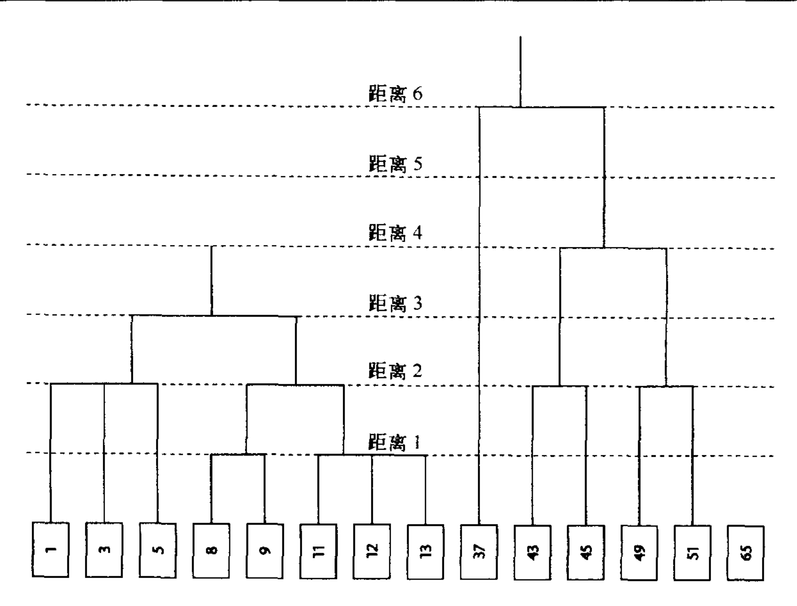


图 11-11 凝聚聚类示例

看到两种方法产生的树之间的相似性，人们自然会问，决策树所使用的算法是否可以用于聚类过程呢？答案是肯定的。决策树算法始于全部记录的集合，然后寻找一种方法把该集合分成不同的纯度更高的分类，它在某种意义上是由一个纯度函数确定的，在标准决策树算法中，这个纯度函数用的是一个单独变量——即目标变量——来做出决策。把决策树变换成聚类算法所需要的就是选取一个纯度函数，用它来最小化簇内的平均距离，或者使簇之间的距离最大化。这样的纯度函数的一个例子就是到上一层簇形心的平均距离。

由于纯度函数没有任何改变，我们可以认为决策树提供定向的聚类，就是说，对于某些目标变量，它们产生具有相似记录的簇。正是因为这一个原因，普通的决策树对于客户分片时常是一个好的选择，远比本章中讨论的非定向聚类要好。如果客户分片的目的是找出那些忠诚的、或者有利可图的、或者有可能响应一些特别的促销服务的客户片断，那就可以把那些变量之一（或一个代表）作为目标变量进行定向聚类；另一方面，如果客户分片的目的是刺激与新产品服务相关的各种自然出现的客户簇的讨论，那么非定向方法更适当。

11.5.4 自组织映像

自组织映像是神经网络的一个变体，很久以来已被应用于二维图像特征探测之类的应用。最近，已被成功用于更广泛的聚类方面。这在第 7 章自组织网络中已经进行了讨论。

11.6 评价簇

当把 K 平均算法用于聚类探测时，是否有方法可以确定什么样的 K 值可以发现最佳簇？同样地，当使用分层方法（hierarchical approach）时，是否有测试可以找出分层的哪一层包含最佳簇？说某个簇好意味着什么？

当在实践中使用聚类的时候，这些问题都是重要的。通常来说，簇中应该有高度相似的成员，或者，从几何学的角度看，这些成员应该彼此很近——而各个簇之间应该分开很远。

衡量簇内相似程度的一个标准度量是方差（簇内的每一个成员与均值之差的平方和），因此，最佳簇可能是有最低方差的那一组。但这个度量没有考虑簇的大小，一个类似的度量可能是平均方差，即总的方差除以簇的大小。

凝聚聚类以方差作为度量就没有意义，因为这个方法总是从一个簇开始，当然，这样方差就是零。对于凝聚簇，一个好的度量是如下两个量之间的差值：形成它的距离值和它合并到下一层的距离值之差，这是对簇持久性的测量。强簇，比如在图 11-11 那个在距离为 3 时把 1 岁与 13 岁相联系的簇，可以在算法中持续多次迭代。

适用于任何形式的聚类探测的通用度量是采用任何的相似性度量或距离度量方法来形成簇，然后使用它来比较同一簇中簇成员和簇形心之间的平均距离与不同簇的簇形心之间的平均距离，这个过程可以单独用于每个簇以及全部簇的集合。

提示： 如果在若干较弱的簇中有一、两个好的簇，可以通过除去所有强簇的成员来改善结果。强簇无论如何都值得进一步分析，而且除去它们的强劲牵引可能在余下的记录中发现新的簇。

11.6.1 在簇内部

聚类过程时常会产生一个或几个强簇——即相当大的簇，其中的记录非常相似。问题是，强簇为什么这么特别？在这些簇的记录中，到底是什么原因导致它们总是集中在一起？甚至更重要的是，由于在该簇中来自数据库的其他部分的噪音已经减少，是否能从中找出可能的规则和模式？

解决这些问题的最容易的方法是计算簇中每个变量的均值，并比较该均值与在初始的总体群体中相同变量的均值，利用差值的大小（或者更好一点，是用 z 得分）把不同变量排序。查看那些具有最大差异的变量有助于解释该簇特殊的原因。

11.6.2 在簇之外

即使只发现一个簇，聚类过程可能也是有用的。当甄别一个非常稀有的缺陷时，可能没有充足的例子训练定向数据挖掘模型来发现它。一个例子是在制造厂测试电动马达。聚类方法只能用于在包含好马达的样本中决定“常态”簇的形状和大小，当一个马达由于任何理由排除在簇之外时，它就是可疑者。这种方法已经被用于医学领域以发现组织中的不正常细胞，以及在无线通讯中发现那些涉嫌欺诈的呼叫模式。

11.7 案例研究：聚类城镇

《波士顿环球报》(Boston Globe) 是服务于波士顿以及东马萨诸塞州和新罕布什尔南部周围区域的两家大日报之一，《波士顿环球报》是波士顿的主流报纸，2003 年的日发行量超过 467 000 份，而《波士顿先驱报》(Boston Herald，该城市的另一份主要日报）的日发行量为 243 000 份。在星期天，环球报的发行量甚至超过 705 000 份。即使处于这样的领先位置，2003 年环球报也不敢有任何懈怠。因为与许多报纸一样，它也面临着这样一些问题：波士顿核心市场读者群在缩减，郊区报业市场面临来自地方报纸的强有力竞争，在那里的一些读者已经流失。

为了与郊区报纸更好地竞争，环球报加入了为不同地区定制的报纸版面，为按照地域划

分的 12 个地区加入了特别编辑内容。每周有两天，读者都可以读到为本区精心整理的一些地方报导页。环球报使用的编辑区域利用的是环球报已有的数据、常识性内容以及地图，但没有正式的统计分析。在编辑区域组成方面有一些限制条件：

- 地域必须是地理上连续的，以便波士顿中心印刷厂运载本地版面的卡车可以选择合理的运输路线。
- 地域必须适度紧凑，且包含足够的人口以证明特殊化编辑的内容是恰当的。
- 编辑区域必须接近于过去做广告的地理地域。

在这些限制条件框架内，环球报希望设计出能够把相似的城镇聚集在一起的编辑区域。这听起来是可行的，但实际上哪些城镇是相似的？这就是《波士顿环球报》给我们这些数据挖掘者带来的问题。

11.7.1 创造城镇特征

在决定哪一些城镇可以归于一起之前，必须找到描述城镇的方法——城镇特征，它需要包括可以用于表征城镇特点，以及可用于比较该城镇及其邻近城镇的每个特征的一个列。碰巧，在数据挖掘者先前研究的一个早期项目“找出增加未来日发行量的潜在城镇”中，已经定义了城镇特征标识。为预测环球报家庭递送穿透度（penetration）而开发的回归模型中使用过的那些客户特征标识（customer signature），对非定向聚类也被证明同样有用。通常会出现的情形是，已经收集的一组有用的描述性属性可以用于所有其他的事情。在另外一个例子中，一家长途公司为了要预测欺骗，开发出了以呼叫明细数据为基础的客户特征标识，后来发现当区别商务和住宅用户时，相同的变量也是有用的。

提示： 虽然产生好的客户特征所花费的时间和精力让人有些畏惧，但这种努力从长期来看是有回报的，因为同样的属性对于许多不同的目标变量经常被证明是有预言性的。这样看来，被经常引用的一个经验之谈，即“数据挖掘项目上花费的时间有80%用在数据准备上”可能就不是那么回事了，因为这种数据准备工作可以在几次预言性模型的建立工作中被分期偿还。

数据

城镇特征标识可以有几个来源，大部分变量可以从 1990 年和 2001 年城镇级的美国人口普查数据（census data）得到。人口普查数据可以提供年龄、种族、宗教信仰、职业、收入、住宅价值、平均通勤时间以及诸多其他令人感兴趣的变量。除此之外，环球报还有外围数据供应商提供的关于订户家庭层次的数据，当然还有每个城镇的发行量数据，以及订阅者层次的信息，如优惠计划、投诉电话和订户类型（日常、周日或两者都是）等。

可以通过四个基本步骤来创建城镇特征标识：

1）聚集。

2）归一化。

3）计算趋势。

4）创建衍生变量（derived variable）。

把这种数据转变为城镇特征标识的第一步，是聚集城镇层次的每种数据。举例来说，聚集订户的数据以得出每个城镇中订户的总数和中值订户家庭收入。

下一步是把计数转变成百分比。大部分人口统计学的信息是以计数形式出现的，甚至像

收入、住宅价值和孩子数目等，也是以预先定义的人均计数来报告的。把所有计数转换为城镇人口的百分比是把人口差别很大的不同城镇的数据归一化的一个好例子。事实是，在2001 年有 4 年大学学历的 27 573 个人住在 Brookline，Massachusetts 的实际情况则没有那么令人感兴趣，它们只相当于教育水平高的城镇的 47.5%，而在波士顿，具有类似学位的人非常多，但只占到那里总人口的 19.4%。

人口普查数据中每个变量都有相隔 11 年的两个值可以使用。这种历史数据是让人感兴趣的，因为由此可以观察趋势。城镇的人口是在增加还是减少？其中，学龄人口有多少？西班牙血统的人口有多少？像这样一些倾向影响了对一个城镇的感觉和印象，因此应该在特征标识中表现出来。对于某些因素，如总人口，绝对的趋势是令人感兴趣的，因此可以计算 2001 年的人口总数与 1990 年的人口总数之比，来表示这种趋势。对于其他一些因素，如城镇中既有租户也有房主，人口中房主比例的改变更有意义，因此可以用 2001 年房主百分比与 1990 年百分比的比率来说明这个问题。在所有的情况中，对于任意随时间增加的量，转换成百分比之后产生的值是一个大于 1 的指数，而对于随时间减小的量，该值是一个小于 1 的指数。

最后，为了取得特征标识中不可辨别的重要城镇属性，可以从已经存在的变量中衍生出另外一些变量。举例来说，离波士顿的距离和方向在形成城镇簇方面似乎看起来很重要，这些数据是以金穹顶州议会大厦的纬度和经度为坐标原点来计算的，Oliver Wendell Holmes 曾经称金穹顶州议会大厦为“太阳系的中心”（今天的波士顿人并不像 Holmes 法官那么谦逊，他们把整个城市称为“宇宙的中心”或只是简单的“中心”［Hub］）。报纸大字标题的作者通常用“hub”代替“Boston”以节省 3 个字母，比如“Hub man killed in NYC terror attack”（一个波士顿男子在纽约市发生的恐怖袭击中遇难）就是一个例子。在线邮政服务数据库给每个城镇经度和纬度提供方便的来源，绝大多数的城镇有单一邮政编码，对于有多个邮政编码的城镇，总是选择以最低数字表示的邮政编码。从某个城镇到波士顿的距离可以容易地从纬度（latitude）和经度（longitude）使用标准的欧几里得几何距离来计算。尽管传说地球是圆的，我们还是用简单的平面几何来进行这些计算：

```
distance = sqrt((hub latitude-town latitude)^2 + (hub longitude-town longitude)^2)
angle = arctan((hub latitude-town latitude)/(hub longitude-town longitude))
```

这些公式是不严谨的，因为它们假设地球是平坦的，且纬度上 1°的长度与经度上 1°的长度相等，不过，我们所要讨论的区域还没有大到让这些“平坦地球假设”出现很大差异的程度。还要提及的是，因为这些数值只是用于彼此相互比较，所以不必把它们转换为我们熟悉的单位，如英里、公里或度等。

11.7.2 创建簇

创建簇的第一步就是利用那些以人口统计学和地理学数据描述该城镇的特征标识，但用这种方法构建的簇不能直接用于创建编辑区域，因为还有地域方面的约束条件，即编辑区域必须由毗邻的城镇构成。由于有相似人口统计学数据的城镇未必是彼此相连，基于城镇特征标识找出的簇包括地图上的所有城镇，如图 11-12 所示。加入权重可以增加形成簇的过程中地理变量的重要性，但结果可能会导致那些非地理变量被完全忽略。因为目标是寻找至少部分地基于人口统计学的相似性，需要更侧重于人口统计学方面，所以就放弃了地域簇的想

法。这样，人口统计学簇就可以与地理学约束因素一样，作为一个因素用于设计编辑区域。

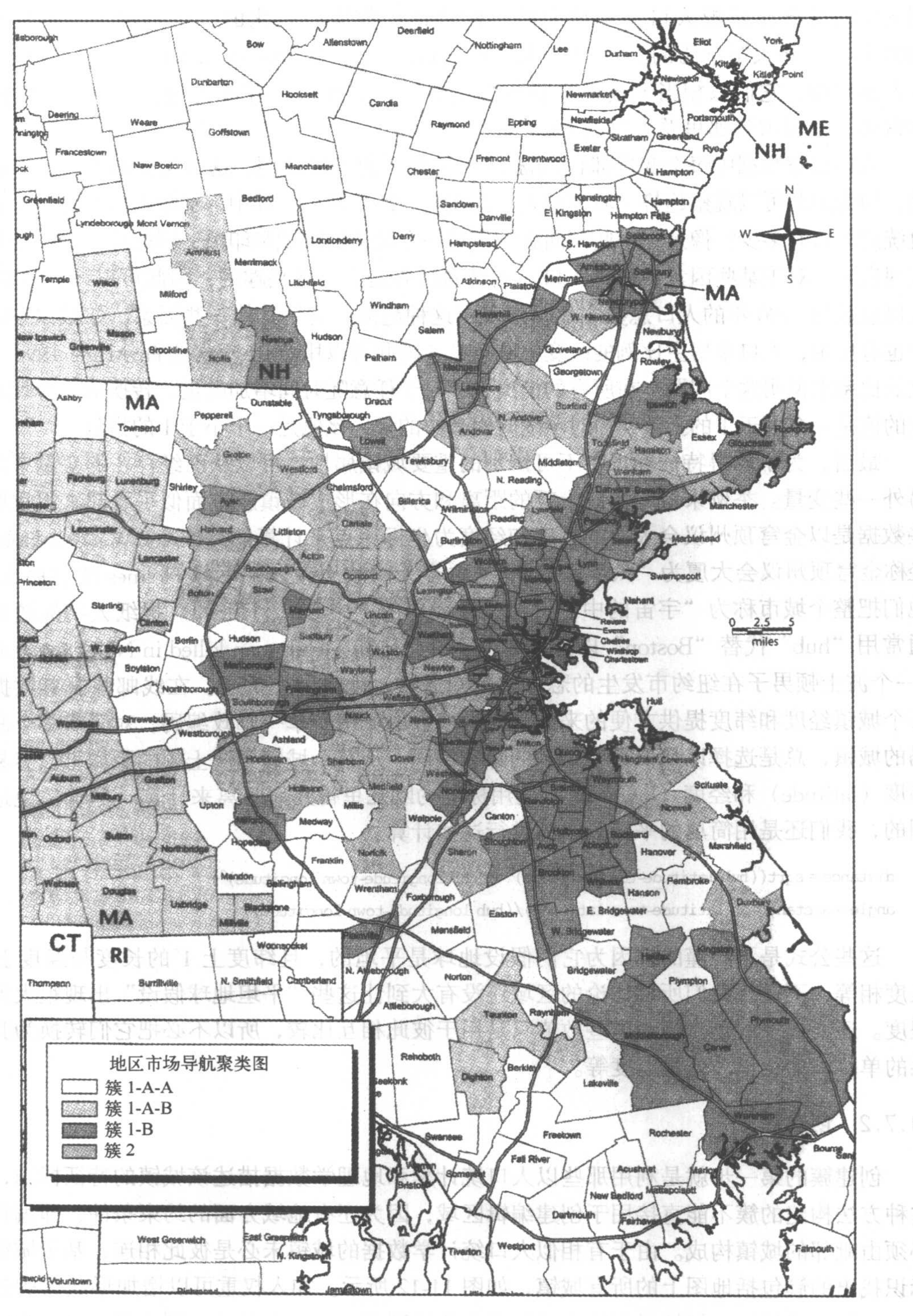


图 11-12 该图显示了 Massachusetts 东部及 New Hampshire 南部各城镇人口统计学聚类情况

确定簇的正确数量

直接通过聚类创建编辑区域的想法遇到的另一个问题是，出于一些商业方面的原因，可能需要 12 个编辑区域，但我们不能保证找到这么多好的簇。这就产生了另外一个问题，即如何为数据集确定合适数目的簇。用于这个聚类过程的数据挖掘工具（MineSet，由 SGI 公司开发，现在可以从 Purple Insight 公司购买），为解决这个问题提供了一种不错的方法，它把 K 平均聚类与分裂树方法结合起来：首先以较低的 K 边界确定簇数目，使用普通的 K 平均算法构建 K 聚类，利用适应度度量（比如不论使用哪一种距离函数都会得到的距簇中心的均值距离或方差）确定哪一个是最差的簇，然后把这个簇分裂为两个簇，反复重复这一过程直到达到某个上界。每一次迭代后，记录该簇集合的总体适应度的度量结果。前面已提及的度量是：从簇成员到簇中心之间的平均距离与簇之间的平均距离之比。

需要记住的是，簇的最重要的适应度度量就是那个难以量化的度量——簇对某个应用的有效性。在图 11-13 所示的聚类树中，聚类树算法的下一次迭代建议把簇 2 进行分裂，形成的簇有明确的差异，但对于任何环球报感兴趣的变量而言，所形成的新簇都没有不同的表现，比如像家庭递送穿透度或订户资历等。图 11-13 显示了最终的聚类树，列出了在叶子上的每个簇的一些统计学数据。

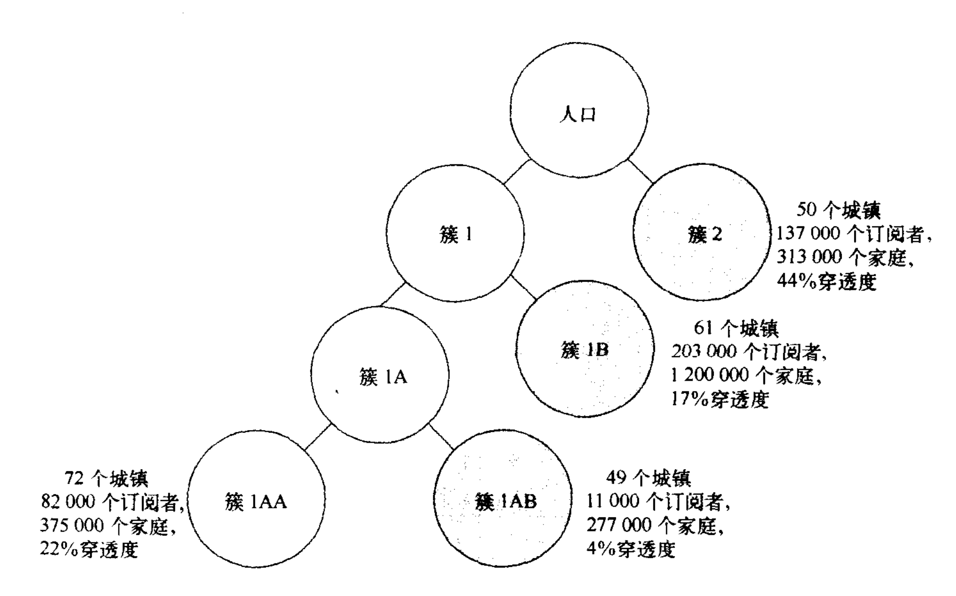


图 11-13　聚类树把《波士顿环球报》所服务的城镇分为四个独特的组

簇 2 包含了 50 个城镇中的 313 000 个家庭，其中 137 000 个家庭订阅日报或周末版环球报，这一家庭递送穿透度使得簇 2 无疑是最佳簇。能够用于把簇 2 与其他簇和人口总体区分开的变量是住宅价值和教育程度。这个簇有最高的住宅价值比例、最高的有 4 年大学学历的人数比例、最高受教育的平均年数和最低的蓝领工作人员比例。从家庭递送穿透度观点来看，次好的簇是簇 1AA，它的一个显著特点就是其平凡性，在这个例子中，住宅价值和家庭收入这两个最重要变量的均值，与总体人口的均值非常接近。簇 1B 的特征是，他们都是一些有最低家庭收入、历时最久且邻近波士顿的订户。簇 1AB 是惟一主要以地域为特征形成的簇，都是远离波士顿的城镇。一点也不令人吃惊的是，其家庭递送穿透度很低。在所有

簇中，簇 1AB 中住宅价值最低，但家庭收入为平均数。可以推断：在簇 1AB 中的人们选择在离城市较远的地方居住，因为他们希望有自己的住宅，而在效区边缘房价比较便宜，这种假设可以在市场调查中得到验证。

11.7.3 利用主题簇调整区域边界

聚类项目的目标是确定已经存在的编辑区域，每个编辑区域都是由处于上述四个簇之一的城镇集合构成，下一个步骤是通过手工方法将某些城镇交换到邻近的区域，以增加每个区域的纯度，例如，表 11-1 中显示：所有处于“City”区域的城镇都位于簇 1B 中，只有 Brookline 除外，它处于簇 2 中。在邻近“West 1”的区域中，除 Waltham 和 Watertown 在簇 1B 中以外，其他所有的城镇都位于簇 2 中。把 Brookline 交换到 West 1 区域，并把 Watertown 和 Waltham 交换到“City”区域，可能会使两个编辑区域纯度增加，因为这样在每个区域中的所有城镇将共享同样的簇赋值，这个新的“West 1”区域会是整个簇 2，而新的“City”区域将是整个簇 1B，就像从图 11-12 中可以看到的，这些新区域仍然是地理上相毗邻的。

有了相似城镇组成的编辑区域，对环球报来说，就可以容易地集中对本地内容提供有针对性的社论，这将带来更高的发行量和更好的广告销售。

表 11-1 “City”和“West 1”编辑区域中的城镇

城 镇	编 辑 区 域	簇分配状况
Brookline	City	2
Boston	City	1 B
Cambridge	City	1 B
Somerville	City	1 B
Needham	West 1	2
Newton	West 1	2
Wellesley	West 1	2
Waltham	West 1	1 B
Weston	West 1	2
Watertown	West 1	1 B

11.8 小结

自动聚类探测是一项非定向数据挖掘技术，可用于认识复杂数据库的结构。通过将复杂数据库分解为简单的簇，自动聚类方法可用于提高更具指导性的那些技术的性能；通过选择不同的距离度量，自动聚类可被应用于几乎任何类型的数据。它可以用于在一系列的新闻题材或保险索赔中发现簇，也同样可容易地用于寻找天文学或金融数据中的簇。

聚类算法依赖于某种相似性的度量，以表明两条记录是相近的还是远离的。通常采用几何学距离的含义，但有另一些可能，当要分类的记录包含非数值型数据时，这些另外的方法可能会更适当。

自动聚类探测最常用的算法之一是 K 平均算法，是一种基于距离寻找 K 聚类的迭代方法。本章还介绍了几种其他聚类法：高斯混合模型是基于 K 平均算法的变体，允许簇之间

的重迭；分裂聚类通过将一个初始的大簇逐次分裂，建立一个聚类树；凝聚聚类则开始于许多小的簇，逐渐把它们结合起来，直到只剩下一个簇为止；分裂算法和凝聚算法都允许数据挖掘者利用外部标准来判断结果聚类树的哪个层次对于某个特定应用是最有用的。

本章介绍了簇适应度的一些技术度量，但聚类过程的最重要度量是，这些簇对于促进某些商业目的到底有多大用处。

第12章　市场营销中的风险函数和生存分析

风险、生存，这些极端术语容易使人联想到可怕的情景：不管是闪烁的蓝光、高尔夫吞球的风险，还是更可怕的事情，比如史蒂芬·金的小说、战争电影或某些真实的电视秀。也许如此可怕的联想说明了为什么这些技术常常不与市场营销联系到一起。

如果真是这样，那是令人惋惜的。生存分析（survival analysis，也称为时间事件分析[time-to-event analysis]）则没有什么好担心的。恰恰相反：生存分析对了解客户非常有用。虽然其根源和术语来自医学研究和制造业的故障分析，但概念是专为市场营销而设计。生存告诉我们何时该开始担忧客户做出的重大决策，如：结束购买关系。它告诉我们哪些因素与事件最紧密相关。风险和生存曲线也提供客户和他们的生存周期的快照，并可以回答这样一些问题，诸如："我们应该在多大程度上担心该客户将要在不久的将来离开？"或"这位客户最近没有进行购买；是否应该开始担忧客户将不再返回？"

生存方法关注客户行为的最重要方面：保有期（tenure）。客户曾经保持多久为我们提供了很多的信息，尤其是当与特别的业务问题相关的时候。客户在未来多长时间将仍然是客户，这是一个谜，但过去的客户行为有助于揭示未来的秘密。几乎每种商业活动都认识到客户忠诚的价值。正如本章后面看到的，一项忠诚度的指导原则——客户停留时间越长，越不可能在任何时间点终止购买关系——确实是一个关于风险的正确表达。

市场营销与医学研究领域有几点不同。其一是，我们的行为结果很少令人感觉可怕：一位病人可能死于拙劣的治疗，然而市场营销的结果仅仅以金钱来度量。另一个重要区别在于数据量。最大的医学研究有数万参与者，并且很多研究结论仅来自其中的数百人。当试图决定平均无故障时间（mean time between failure，MTBF）或平均故障间隔时间（mean time to failure，MTTF）（这是制造业中一个描述某个昂贵的机件直到损坏所需时间的术语）时，结论时常基于几十个故障。

在客户世界中，数万只是一个较低的限度，因为客户数据库时常包含数以百万计客户和前客户的数据。生存分析的许多统计背景集中于在数百数据点中提取每一点信息。在数据挖掘应用中，数据量是如此巨大，以至于人们对置信度和精确度统计的关注被管理大量数据的关注所代替。

生存分析的重要性是它提供了解时间事件特征的方法，如：

- 客户何时可能离开
- 一位客户可能转向新客户片段的未来时间点
- 客户可能拓宽或者缩小客户关系的未来时间点
- 客户关系各种因素中，增加或者减少保有期的因素
- 各种因素对客户保有期的定量影响

这些对客户的深入了解直接馈入市场营销过程中，使得了解不同客户组停留的时间成为可能，进而可以得知从这些客户群组可能赚取的利润。由此可以预测客户数目，可以同时考虑到新客户的获取和当前客户群的下降。生存分析也使得确定哪些因素（包括在客户关系创建之初和后续阶段中的各种因素）对客户停留最长时间、影响最大成为可能。生存分析也可

应用于客户保有期之外的其他事情，可以确定何时另一个事件不再可能发生，比如客户转向一个网站。

开始讨论生存分析的很好起点应该是客户保持的可视化，保持是生存的粗略近似值。在这之后再讨论风险，即生存的组成模块。这反过来要结合生存曲线，生存曲线类似于保持曲线（retention curve)，但更有用。本章最后以 Cox 比例风险回归和生存分析的其他应用结束讨论。在这个过程中，本章提供生存分析在商业环境中的特殊应用。与所有的统计学方法一样，生存有一个深度，这远远超出了本章的内容范畴，本章中我们试图避免这些技巧所涉及的复杂数学内容。

12.1 客户保持

客户保持是大多数商业活动关于其客户的一个常见概念，因此是一个好的讨论起点。保持实际上是非常接近生存的近似值，尤其是在考虑一组客户全部同时开始的时候。保持提供了一个熟悉的框架，用以引入一些重要的生存分析的概念，如客户半衰期和平均截取客户保有期。

12.1.1 计算保持

客户会停留多长时间？这样看似简单的问题在现实世界中变得比较复杂。了解客户保有期需要两方面的信息：

- 每位客户何时开始
- 每位客户何时停止

这两个值的差就是客户保有期，是客户保持的一个很好的度量。

任何合理的客户数据库都应该使这些数据易于使用。当然，市场营销数据库很少是简单的。对这些概念有两种挑战：第一个挑战是确定什么是开始和停止，这个决定通常依赖于商业类型和可用的数据。第二个挑战是技术方面的：在可用的数据中发现这些开始日期和停止日期不像它们最初出现时那么容易。

对订阅和基于账户的商业，开始和停止日期很好理解。客户在某个特定的时间开始订阅杂志，当不想再为杂志付账的时候结束订阅。客户在一个特定的时间在电信服务、银行账户、ISP 服务、电报服务、保险条款或者电力服务等合约上签字，在另一个时间取消服务。所有这些情况下，开始和结束关系是明确定义的。

其他一些商业活动没有这样连续的关系。转账业务尤其如此，如零售、网站门户和目录销售等，每位客户的购买（或者访问）在时间上很分散，或者可能仅仅一次。关系的开始是清楚的，通常情况下是第一次购买或者访问站点。结束则很难判断，有时候是通过商业规则产生的。例如在过去的 12 个月里没有购买的客户，可能视为流失。基于这些定义，客户保持分析可以产生很多有用的结果，类似的应用领域是确定一个时间点，在这个时间点之后，客户不可能返回（本章后面有一个这样的例子）。

技术方面的问题更具有挑战性。我们来考虑杂志订阅问题：客户关系从客户签订订单的日期开始吗？抑或是从第一次收到杂志开始？那时可能已经是几周之后了。或者，客户关系的开始是在促销期结束且客户开始支付的时候？

尽管所有这三个问题都是客户关系的重要方面，但是焦点应该是客户关系的经济方面。

成本和（或）收入在账户启动时开始使用（即订阅杂志的开始日期），在账号停止时结束。为了解客户，除了开始订阅的日期之外（平时签合同的客户与周末签合同的客户是否不同?），合同的日期和时间无疑是重要的，但是，这不是经济关系的开始。而促销阶段结束时确实是客户关系的*初始条件或者零时协同变量*。当客户签订合同时，初始促销阶段是已知的。生存分析可以从这些初始条件收益来精修模型。

12.1.2 保持曲线揭示的内容

一旦可以计算保有期，就可以将其体现在*保持曲线*图上，从而可以显示在一段特定的时间保持的客户比例。这实际上是一种累积直方图，因为有三个月保有期的客户被包含在1个月和2个月的部分中。因此，保持曲线总是从100%开始。

现在，我们假设所有客户从同一时间开始。例如，图12-1比较了10年以前在同一个时间点开始的两组客户的保持。曲线上的点显示了在1年、2年处保持的客户比例，依此类推。这样的曲线从100%开始，然后逐渐下降。当保持曲线表示的客户在同一时间开始的时候（正如这一情形），它就是生存曲线的一个接近的近似值。

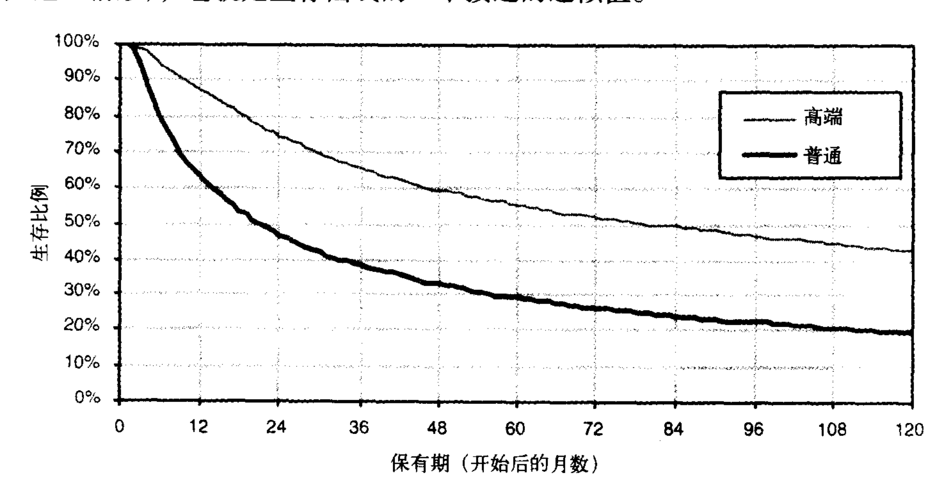


图12-1 保持曲线表明，高端客户停留时间更长

不同的客户组之间保持的差别在图中清晰可见，并且可以量化。最简单的度量是考察特定时间点的保持。例如，10年以后，普通客户的24%仍然处于客户圈，而且这些客户中只有1/3的人持续了5年。高端客户做得更好，其中超过一半的人持续了5年，42%的人其客户生存期至少是10年。

比较不同组的另一种方法是确定一半客户离开的时间，即客户的半衰期（统计学术语是*中值客户生存期*）。因为极少数具有很长生存期和很短生存期的客户不影响中值，所以这是一个很好的度量。一般来说，中值对少数的离群值不敏感。

图12-2阐明如何使用保持曲线找到客户半衰期（customer half-life），即恰好保持50%的客户的点，也即水平格线的50%处和保持曲线的交点。这两个组的客户半衰期展示了与10年生存分析完全不同的差别：高端客户的中值生存期接近7年，而普通客户的中值生存期略低于2年。

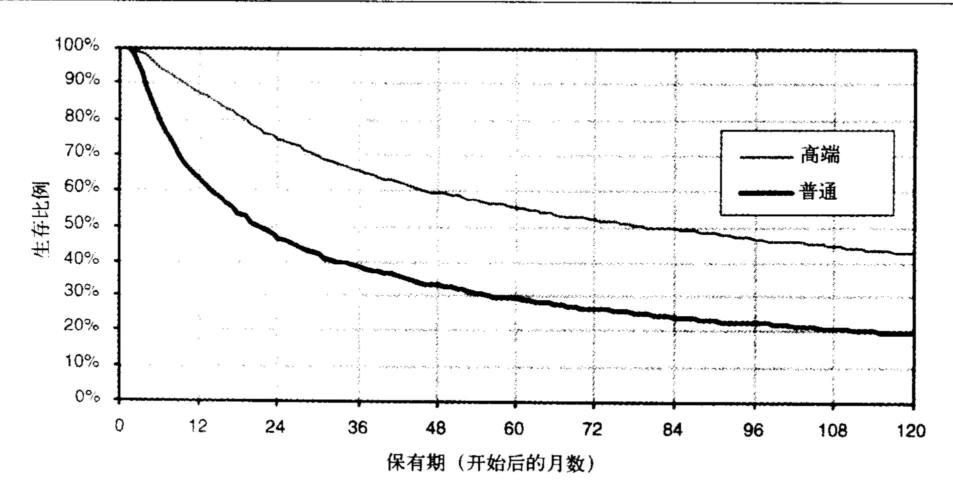


图 12-2　中值客户生存期是保持曲线与 50%处相交的点

12.1.3　从保持曲线找出平均保有期

客户半衰期对于比较很有用且容易计算，因此是一个有用的工具。然而，它不回答这样一个重要的问题："一般说来，多少客户在这期间是有价值的?"回答这个问题需要有每个时间的平均客户价值和所有客户的平均保持程度。中值不能提供这些信息，因为中值仅仅描述恰好为一半的客户的情况，也即在 50%等级的那些客户。关于平均客户价值的问题需要估计所有客户的平均剩余生存期。

计算平均剩余生存期的一个简单的方法是：在此期间的平均客户生存期就是保持曲线下方的面积。图 12-3 用一种巧妙的可视方法展示了该计算。

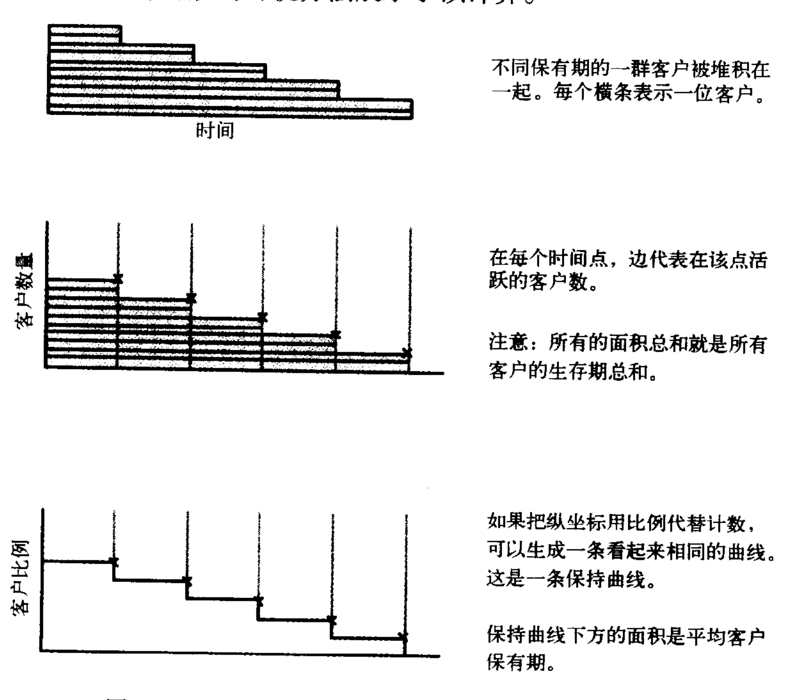


图 12-3　根据保持曲线下方的面积计算平均客户保有期

首先，设想客户全部躺下，脚在左边排成一排。他们的头部表示其保有期，因此就可以用各种不同身高（或者体宽，因为他们是水平的）的客户代表各种不同保有期的客户。为可视化起见，保有期较长的客户躺在底部，支撑起较短保有期的客户。连接他们的鼻子的线可看做在特定时段保有的客户数目（记住：假定所有的客户是在同一时间点开始）。这条曲线下方的面积是所有客户的保有期总和，因为水平躺着的每位客户都被计算在内。

用垂直轴除以总的计数会产生一条保持曲线，用百分比代替了计数。在曲线下方的面积是总保有期除以客户计数——瞧，特定时期的平均客户保有期就体现在图中。

提示：在客户保持曲线下方的面积是在曲线上一段时间内的平均客户生存期。例如，对于一个有 2 年数据的保持曲线，曲线下方的面积就代表两年的平均客户保有期。

这一简单的观察解释了如何获得平均客户生存期的估算值，这样当一些客户仍然是活跃的时候，就提供一个说明。平均值实际上是在保持曲线之下时间段的平均。

考虑本章较早提到的保持曲线。这些保持曲线是关于 10 年时间的，因此在曲线下方的面积是最初 10 年期间客户关系的平均客户生存期估计。对于 10 年后仍然活跃的客户，没有方法知道他们是否会在 10 年后的第一天全部离开；或者他们是否全部都将再保持 100 年的时间。由于这个原因，在所有的客户已经离开之前，不可能确定实际的平均保持时间。

这个值非常有用，它被统计学家称为*截取均值生存期*。如图 12-4 所示，较好的客户 10 年的平均生存期是 6.1 年；其他组的平均生存期是 3.7 年。一般说来，如果一位客户的价值是每年 \$100 的话，那么开始之后的 10 年间，高级客户的价值超过一般客户的价值 \$610 − \$370 = \$240，或大约每年 \$24。这 \$24 可能是特别为高级客户设计的保持计划的利润，或者可能给出这样的保持计划的预算数目的一个上限。

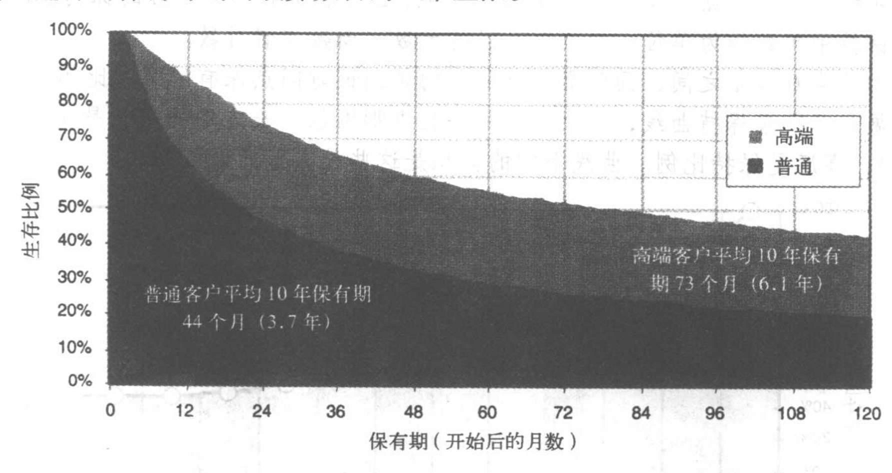


图 12-4 不同客户组的平均客户生存期可以利用保持曲线之下的面积进行比较

12.1.4 把客户保持看做衰变

虽然我们通常不主张把客户和放射性材料作类比，但是这一类比有助于理解保持。把客户喻作铀块，它慢慢地、放射性地衰变为铅。好客户是铀，已经离开的客户是铅。随着时间的过去，留在块中的铀数量看起来有点像我们的保持曲线，微妙的区别在于铀的时间帧以十亿年计，而不是较小的时间度量。

铀的一个非常有用的特征是已知的，或更准确地说，科学家已经确定如何计算特定的时间之后恰好还剩多少铀。之所以能够这样做，是因为他们已经创建了描述放射性衰变的数学模型，而且这些已经被实验证实。

放射性材料的衰变过程被描述为指数衰变（exponential decay)。这意味着不管时间过去多长，都有相同比例的铀变成铅。例如，最常见的铀，其半衰期大约为 45 亿年，所以大约一半的铀在这一时间后会变为铅。在后续的 45 亿年之后，剩余的铀的一半将会衰变，最后剩下最初的 1/4 为铀，3/4 为铅。

警告：指数衰变有许多有用的性质，可用于预测观察范围之外的事情。不幸的是，客户几乎不表现为指数衰变。

指数衰变如此有用的原因是衰变适合一个精确的简单方程，利用这个方程，可以确定在任何给定时间还剩余多少铀。假如客户保持有一个这样的方程岂不是很好?

这当然会非常好，但是不太可能，正如后面部分所示的例子一样："参数方法不起作用"。

为了弄明白这一问题，假想有这样一个世界，其中的客户确实具有指数衰变特性。为便于讨论，这些客户的半衰期假设为 1 年。在一个特定的日期开始的 100 位客户，恰好 50 位在 1 年后仍然活跃。2 年之后，25 位活跃，75 位已经停止。指数衰变可以很容易地预测未来活跃客户的数目。

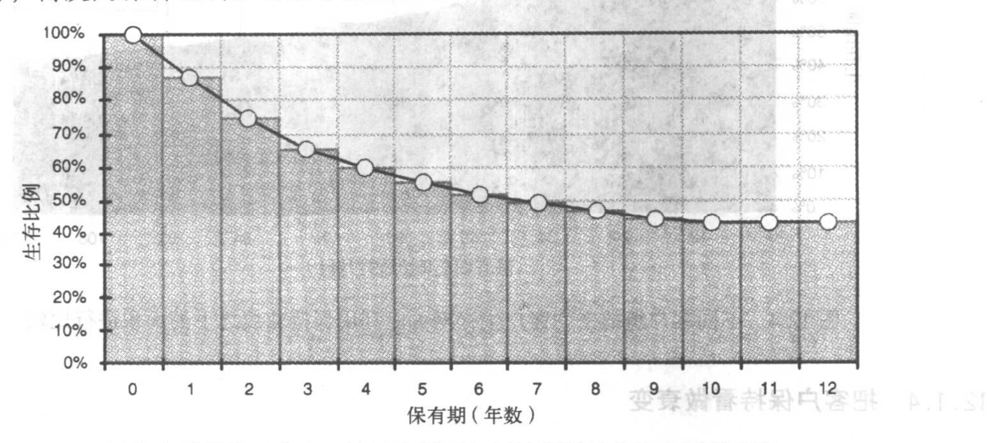
确定保持曲线下方的面积

确定在保持曲线下方的面积看起来像一个使人畏缩的数学运算，但幸运的是，根本不是这么回事。

保持曲线由一系列点构成；每个点代表 1 年、2 年、3 年……之后的保持情况。在这种情况下，保持用年数作为单位；单位也可以是天数、周数或者月数。

每点的值在 0 和 1 之间，因为那一点代表的是到该时间点保有的客户比例。

下图展示了一条保持曲线，每个点用一个长方形围住。长方形的底长是 1（用横坐标的单位测量)，高度是保持比例。曲线下方的面积是这些长方形的面积总和。



用长方形圈住每个点，清楚表明了如何计算保持曲线之下的面积

每个长方形的面积是底乘以高，刚好是保持的比例。那么，所有长方形的总和就是曲线上所有保持值的总和，这在电子数据表中是很容易计算的。瞧，非常简单的一个计算面积的方法，也是相当有意义的观测结果：保持值的总和（按百分比）就是平均客户生存期。我们

还注意到，每个长方形的宽度是一个时间单位，不管横坐标的单位是什么。因此，平均值的单位也就采用横坐标的单位。

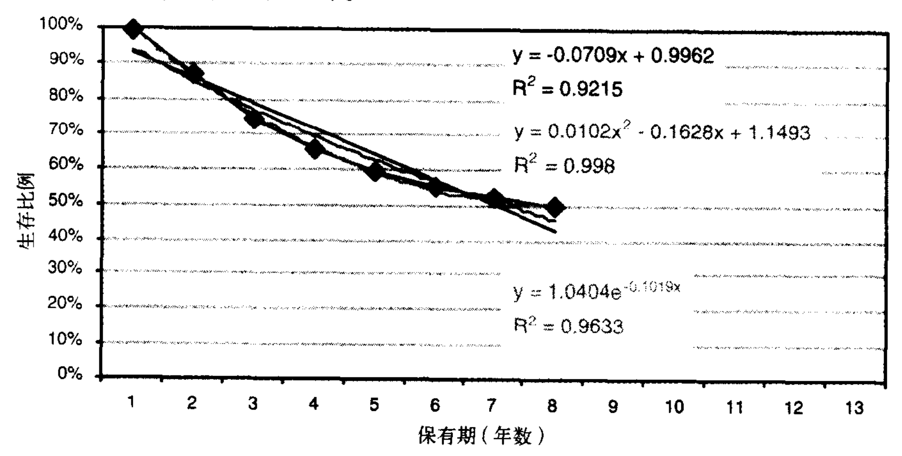
参数方法不起作用

尝试用一些已知的函数拟合保持曲线是非常吸引人的，这种途径被称为参数统计学，因为它利用一些参数描述函数的形状。这种方法的作用是我们可以用它来估计未来发生的事情。

对这种函数来说，直线是最常见的形状。一条直线有两个参数，即直线的斜率和它与Y轴的交点位置。另一种常见的形状是抛物线，还包含一个 X^2 项，因此抛物线有三个参数。描述放射性衰变的指数实际上只有一个参数，那就是半衰期。

下图是关于最初7年的数据的局部保持曲线。

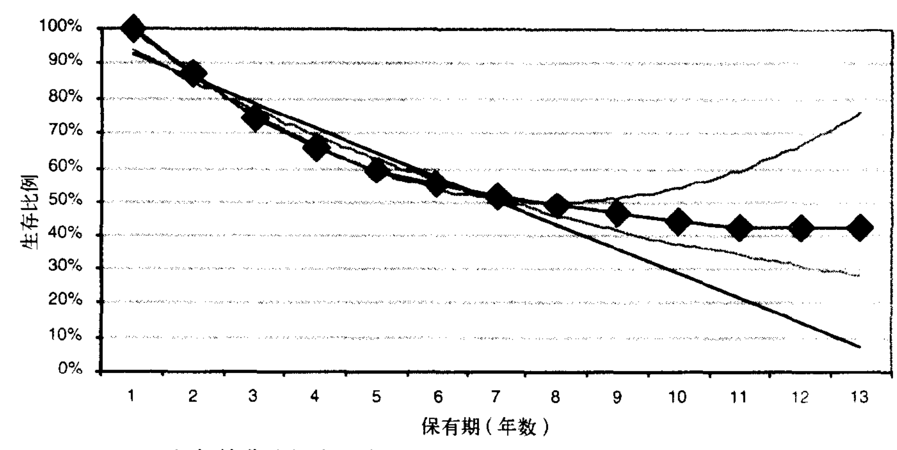
该图也表明三条最佳拟合曲线。注意，所有这些曲线都很好地与这些数值吻合，吻合程度的统计学度量是 R^2，其变化从0到1。超过0.9的值是非常好的，因此按照标准统计度量标准，所有这些曲线都拟合得相当好。



把参数曲线拟合到保持曲线是很容易的

真正的问题不是这些曲线对定义范围内的数据吻合得有多好，我们想知道的是，这些曲线在最初53周之外的效果有多好。

下图回答了这个问题，它向前外推了另外5年的曲线，曲线很快偏离实际数值，而且看来我们走的越远偏离似乎增加得越快。



与保持曲线拟合的参数曲线在定义范围之外拟合得不好

当然，这个图并不证明参数方法不起作用。也许存在这样一些函数，如果具有正确的参数，可能会与观察到的保持曲线拟合得很好，而且在用于定义参数的范围之外，仍然能够继续使用。然而，这个例子确实表明了使用参数方法直接逼近生存曲线所面临的挑战，即使使用更多的数据点，这也与我们的经验一致。能够很好地拟合保持曲线的那些函数，结果也将很快偏离。

描述这种情况的另一种方法是，保有 1 年的客户行为就像新客户一样。考虑一组包含各种不同保有期的 100 位客户，50 位客户后来离开了，不管那一年初期的客户保有期如何——指数衰变是说一半客户将要离开，不管他们的初始保有期如何，这就意味着保有一段时间的客户不如比较新的客户忠诚。但是，通常的情形是保有一段时间的客户实际上是比新客户更好。无论如何，较长保有期的客户在过去已经逗留，并且在未来或许比新客户更不可能离开。指数衰变是一种差的情形，因为它进行了相反的假设：客户关系的保有期对客户离开的比率没有影响（最坏的情况是具有较长期限的客户离开的比率一贯比新客户高，即所谓的“熟而无礼”情形）。

12.2 风险

关于保持曲线的前述讨论表明保持曲线是多么有用。这些曲线很容易理解，但是只有在与相关的数据对应的时候。它们没有通用的形状，没有参数形式，也没有关于客户衰变的重要理论，数据本身就是信息。

风险概率扩展了这个理念。正如这里讨论的，它们是非参数统计方法的一个例子——让数据说明事实，而不是找出一个特别的函数说明它。完全依赖经验的风险概率只是让历史数据决定可能发生的事情，并不尝试拟合数据到某种预想的形式。它们也提供对客户保持的某种估计，可能生成一条精修的保持曲线，我们称之为生存曲线。

12.2.1 基本思想

风险概率回答下列问题：

假设一位客户已经保有一段特定长的时间，因此客户保有期为 t。那么客户在时刻 $t+1$ 之前离开的概率是多少？

描述这个问题的另一个方法是：在时刻 t 的风险就是在时刻 t 和时刻 $t+1$ 之间损失客户的危险程度。当更详细地讨论风险时，使用这一定义可能更有用。对于许多类似的简单理念，风险具有重要的地位。

为提供一个风险的例子，让我们暂时步出商业界，来考虑寿命表，它可以描述某人死于一个特定年龄的概率。表 12-1 展示了 2000 年美国人的这一数据。

表 12-1 以寿命表展示的美国 2000 年的死亡率风险

年　龄	每一个年龄段中死亡人数占总人口的百分比
0～1 岁	0.73%
1～4 岁	0.03%
5～9 岁	0.02%
10～14 岁	0.02%

（续）

年 龄	每一个年龄段中死亡人数占总人口的百分比
15～19 岁	0.07%
20～24 岁	0.10%
25～29 岁	0.10%
30～34 岁	0.12%
35～39 岁	0.16%
40～44 岁	0.24%
45～49 岁	0.36%
50～54 岁	0.52%
55～59 岁	0.80%
60～64 岁	1.26%
65～69 岁	1.93%
70～74 岁	2.97%
75～79 岁	4.56%
80～84 岁	7.40%
85～89 岁	15.32%

寿命表是一个关于风险的好例子。每 137 个婴儿中有 1 个在 1 岁之前可能死亡（这实际上是一个非常正常的比率；在不发达的国家这个比率可能高很多倍）；然后死亡率骤然跌落，但最终稳定地升高。一直到 55 岁的时候，死亡的风险又像 1 岁时那样高。这是某些风险函数的特征形状，被称为*浴缸形*（bathtub shape）。开始时风险很高，很长一段时间保持低水平，然后又逐渐增加。图 12-5 用这一数据为例说明浴缸形特征。

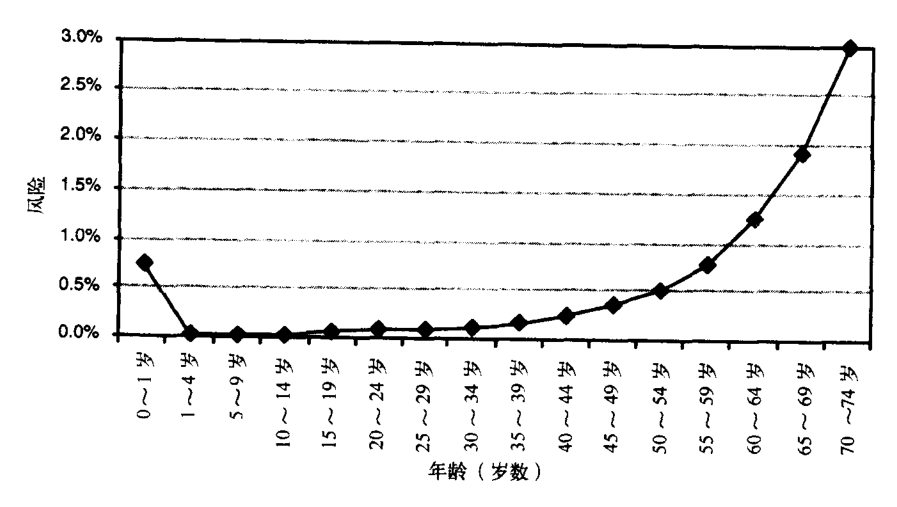


图 12-5 浴缸形风险函数，开始时高，骤然下降，然后又逐渐增加

同样的理念可以用于客户保有期，尽管客户风险更典型的是以天、周或月计而不是以年计。对于给定的保有期 t 计算风险，只需要两笔数据。第一是在时刻 t（或在 t 和 $t+1$ 之间）停止的客户数；第二是可能在这段时期停止的客户（也称为*风险人群*）总数，这包括保有期大于等于 t 的所有客户，也包括 t 时刻停止的那些人。风险概率是这两个数字的比，并

且作为概率，风险总是介于0和1之间。这些风险计算是由统计软件SAS和SPSS中的寿命表函数提供，当然也可以在数据表中进行计算，直接使用客户数据库中的数据。

一条防止误解的说明：为使计算精确，包含在人口计数中的每位客户必须在该特定时间内有停止的机会。这是用于计算风险的数据具有的性质，而不是计算的方法。在多数情况下，这不成问题，因为风险是由全体客户或基于初始条件（如初始产品或商业活动）的某个子集来计算的。当客户包括在计算客户保有期的人群计数中时，不会有任何问题，客户可以在那个时刻之前的任一天停止且仍然被包含在数据集当中。

注意：不要选取在过去某段时间（如过去的一年）内已经停止的客户子集。这样做有什么问题呢？考虑保有期2年且在昨天停止的一位客户的情况：这位客户包含在计算第一年风险的所有人口计数中，但该客户不可能在保有期的第一年期间停止。停止可能发生在过去的一年多之后，这就把该客户排除在数据集之外。因为把不可能停止的客户包含在人口计数中，人口计数太大，以至于初始风险太低。本章后面将介绍一个解决这个问题的方法。

警告： *为了得到精确的风险和生存曲线，使用仅仅基于初始条件定义的客户群。特别地，不要基于客户如何和何时离开来定义群。*

当人口基数很大的时候，没有必要担忧诸如置信度和标准误差的统计学概念。然而，当人口基数很小的时候，就像在医学研究或某些商业应用中一样，置信区间就可能变成一个议题，这意味着比方说5%的风险可能实际上是介于4%和6%之间。在处理小的人口基数（如少于数千人）情况的时候，使用能够提供标准误差信息的统计学方法可能是一个好主意。当然，对于大多数的应用来说，这不是关注的重点。

12.2.2 风险函数示例

讲到这里，我们有必要看一些风险函数的例子。这些例子的目的在于通过考察风险概率，帮助了解发生的事情。前两个例子是基本的，而且事实上本章中已经介绍过。第三个来自现实世界的数据，而且是一种很好的体验，体会风险如何用于提供客户生存期的X光片。

1. 恒定的风险

恒定的风险几乎不需要用图解释。它指的是客户离开的风险完全相同，不管已经保有客户多长时间。这看来像图中的一条水平线。

假如风险以天计，而且是一个常数0.1%，那么每天在1000位客户中有一位离开。这意味着，在一年（或者365天）之后大约30.6%的客户已经离开，一半客户离开大约需要692天，再需要另外的692天使其中一半离开，如此继续。

恒定的风险意味着客户离开的机会不随客户保有的时间长度而变化。这听起来很像指数保持曲线，即放射性元素衰变那样。事实上，恒定的保持风险会使保持曲线遵照一种指数形式。之所以说“会”，仅仅因为尽管在物理学方面确实发生过，但是在市场营销方面不常发生。

2. 浴缸形风险

美国人口寿命表提供了浴缸形风险函数的一个例子，在生命科学中这很普遍，尽管浴缸形曲线出现在其他的领域。就像先前提到的，浴缸形风险开始时相当高，然后下降并且很长的一段时间转为水平，最后风险再次增加。

导致出现这种情况的一种现象是，客户已经签订合同（例如移动电话或ISP服务），典

型的是 1 年或更长的时期。在合同的初期，客户停止是因为服务不合适或者因为不支付账单。在合同期间，客户被阻止取消服务，要么是因为经济上处罚的威胁，要么仅仅是由于感到有责任去遵从初始条款。当合同到期时，客户时常迅速离开，而且高的离去率会持续一段时间，因为客户已经从合同中解放出来。

一旦合同过期，可能有其他一些理由导致客户离开，如产品或服务价格不再有竞争力。市场改变了，客户就会回应这些变化。当电话费下降时，客户更可能流失到竞争对手一方，而不会与当前的服务提供商商讨降低费率。

3. **真实的例子**

图 12-6 展示了一个真实的风险函数的例子。在一家销售基于订阅服务（具体的服务并不重要）的公司，风险函数测量客户在注册后给定的周数停止的概率。

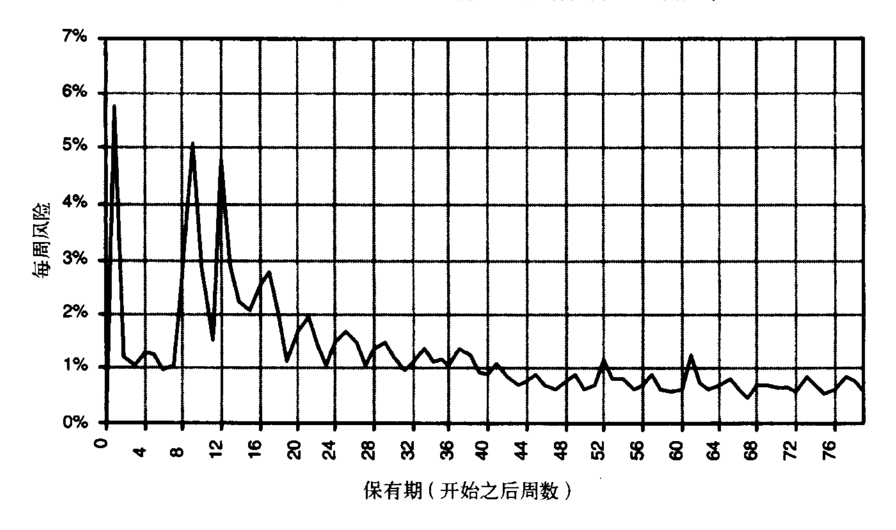


图 12-6　订阅业务的客户风险概率看起来像这样

该曲线有一些值得关注的特征。首先，它开始时高。这是那些注册的客户，但是由于某种技术原因不能开始，诸如信用卡尚未被认可。在某些情形下，客户没有认识到他们已经正式注册——这是在打往外地的电话销售活动中作者最经常遇到的一个问题。

其次，有一个 M 形特征，大约在 9 周和 11 周有两个峰值。第一个峰值，大约在 2 个月，因为没有支付产生第一个峰值。从来不支付账单的客户，或信用卡记账被取消的客户，由于没有支付，大约 2 个月之后停止。由于显著数量的客户在这个时间离开，风险概率很高。

"M"的第二个峰值与提供优惠价格的初始促销活动的结束相一致。典型地，这种促销持续大约 3 个月之久，然后客户必须开始支付全价。很多人决定他们不再需要这项服务。极有可能的是，这些客户中的很多人又从其他促销活动中受益，这是一个与风险相关讨论关系并不密切，但与商业有关的重要事实。

在最初的 3 个月之后，风险函数不再有真正的峰值。大约每 4 或 5 周，有一个小的周期性峰值，这符合每月付账的周期，客户可能是在收到账单之后停止的。

该图也表明风险率呈缓和下降趋势。这种下降是一件好事，因为客户保有期越长，客户越不可能离开。对于这种现象的另一种表达方法是：客户在公司停留的时间越长，忠诚度越高。

12.2.3 审查

迄今为止，对风险的介绍掩盖了生存分析的最重要的概念之一：审查（censoring）。记住风险概率的定义是，在给定的时刻 t 停止的客户数除以该时刻客户总数。显然，如果一位客户在时刻 t 之前已经停止，那么该客户不包含在人口计数中，这是最基本的审查例子。已经停止的客户不包括在他们停止后的计算中。

还有另一个审查的例子，它有些微妙。考虑保有期为 t 但当前仍然活动的客户，这些客户不包括在保有期 t 的风险人口中，因为客户可能仍然在 $t+1$ 之前停止——今天在，明天离开。这些客户已经被排除在特定风险计数之外，尽管他们被包含在 t 为较小值的风险计算中。审查——从某种风险计算中去除一些客户——被证明是一项强有力的技术，对很多生存分析至关重要。

让我们用图来解释这一点。图 12-7 展示了一组客户及其客户关系的开始和结束。特别地，结束用一个空心或者实心的小圆圈表示。当圆圈是空心的时候，客户已经离开，而且他们的准确保有期是已知的——因为停止日期已知。

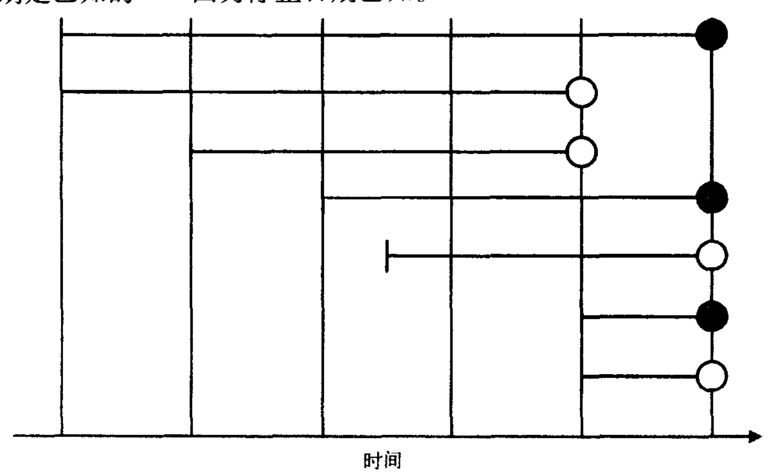


图 12-7 开始于不同时间的客户组，一些客户被审查，因为他们仍然是活跃的

实心的圆圈意味着，客户已经保持到分析日期，因此停止日期尚不知道。该客户——或者特别是该客户的保有期——被审查。保有期至少是当前的保有期，但是可能更大。到底多大是未知的，因为该客户的确切停止日期还没出现。

下面略过这些客户的风险计算，来特别关注审查的任务。在考察客户数据进行风险计算时，需要保有期和审查标记。对于图 12-7 的客户，表 12-2 给出了这些数据。

表 12-2 几位客户的保有期数据

客户	审查	保有期
2	N	4
3	N	3
4	Y	3
5	N	2
6	Y	1
7	N	1

考察在每个时间段发生的事情具有指导意义。在任何时间点，客户可能处于三种状态之一：活跃，即关系在继续；停止，即客户在那段时间停止；审查，即客户不包含在计算之中。表 12-3 展示在每个时间段内发生的事情。

表 12-3 在几个时间段追踪客户

客户	审查	客户寿命	时刻 0	时刻 1	时刻 2	时刻 3	时刻 4	时刻 5
1	Y	5	活跃	活跃	活跃	活跃	活跃	活跃
2	N	4	活跃	活跃	活跃	活跃	停止	审查
3	N	3	活跃	活跃	活跃	停止	审查	审查
4	Y	3	活跃	活跃	活跃	活跃	审查	审查
5	N	2	活跃	活跃	停止	审查	审查	审查
6	Y	1	活跃	活跃	审查	审查	审查	审查
7	N	1	活跃	停止	审查	审查	审查	审查

注意在表 12-4 中，审查的发生时间比生存期晚一个时间单元。即客户 ♯1 生存至时刻 5，之后发生的事情是未知的。给定时间的风险是停止的客户数除以包括活跃的和停止的所有客户总数。

表 12-4 从时间到风险

	时刻 0	时刻 1	时刻 2	时刻 3	时刻 4	时刻 5
活跃	7	6	4	3	1	1
停止	0	1	1	1	1	0
审查	0	0	2	3	5	5
风险	0%	14%	20%	25%	50%	0%

时刻 1 的风险是 14%，因为 7 位客户中有一位在这一时刻停止，所有 7 位客户都保持到时刻 1，尽管都有可能停止，而这里，只有一位停止。在时刻 2，5 位客户留了下来——客户 ♯7 已经停止，客户 ♯6 被审查。这 5 位中有一位停止，风险是 20%，其余的不再赘述。本例说明当考虑到一些（希望许多）客户尚未停止的事实时，如何计算风险函数。

这一计算也表明风险高低是不稳定的——在最后 3 天内从 25% 跳跃到 50% 又到 0%。通常情况下，风险变化不会这么大。这种不稳定，仅仅是因为这个简单的例子中包含的客户少。同样地，在表中对客户排队，是为了便于展示在可管理的数据集上进行的计算。在真实的世界中，这样的方法是不可行的，因为可能有数千或者数百万客户要记录，并且保有期可能是数百或数千天。

另外值得一提的是，这种对风险的处理是把它们作为条件概率来介绍的，其值介于 0 和 1 之间。这种情况是可能的，因为风险使用像天数或周数这样的不连续的时间单元，这是一种可用于客户相关分析的时间描述。然而，统计学家时常采用风险率代替概率，这两个概念关系显然非常密切，但是使用比率的数学方法中包括使人畏惧的积分和复杂的指数函数，并且很难解释对于这个或者那个因素的调整。针对我们的目的，简单的风险概率不但比较容易解释，而且也能解决利用客户数据进行工作时出现的问题。

12.2.4 其他类型的审查

前面的小节介绍两种情况下的审查：客户停止后的风险和仍然活跃的客户风险。还有一

些其他有用的情形。为解释其他类型的审查，请返回到医学领域。

假设你是癌症疾病研究人员，而且已经发现治疗癌症的一种药物。这时你必须开展一项研究，验证这些新治疗药物是否起作用。这类研究通常要追踪一组病人治疗之后的几年时间，如 5 年。对于本例，只需要知道病人是否在研究期间死于癌症（医学研究还有其他需要关注的情况，如疾病的复发，但在这个简单的例子中，我们不考虑这类问题）。

因此你找出 100 位病人，给予相应的治疗，而且他们的癌症似乎已经治愈。你追踪他们几年时间，这期间，七位病人通过去冰岛旅游来庆祝新生。在一件可怕的悲剧中，所有的七位病人碰巧死于由水下火山所导致的一次雪崩。你的治疗对癌症死亡率的有效性是多少？仅看数据而言，似乎有一个 7%的死亡率。然而，这个死亡率显然与治疗无关，因此感觉结论并不正确。

事实上，答案的确不正确。这是风险竞争的一个例子。研究对象可能活着，也可能死于癌症，或可能死于遥远的岛上发生的登山意外事故。或者该病人可能移居塔希提岛而脱离研究，正如医学研究人员所说，这位病人已经“没有必要追查到底”。

解决的方法是审查在被研究事件发生之前退出研究的病人。如果病人退出研究，那么到他们离开的那一时刻之前病人仍然是健康的，在这段时间获取的信息能用来计算风险，而此后无法知道发生的事情。它们在退出的时间被审查，如果一位病人死于其他情形，他或她在死亡的那一时刻被审查，而且该死亡不包含在风险计算中。

提示：处理竞争风险的正确方法是对每种风险设置不同的风险组，其他的风险被审查。

在商业环境中，竞争风险也非常常见。例如，时常有两种类型的停止：客户决定离开的自发停止和公司决定该客户应该离开的强制停止——时常是由于没有支付账单。

在分析自发流失（voluntary churn）时，对由于不支付账单被迫停止客户关系的客户，会出现什么情况？如果这样的一位客户被迫在第 100 天停止，那么该客户在 1 至 99 天就没有自发地停止，这些数据能够用于生成自发停止的风险。然而，从第 100 天开始，客户被审查，如图 12-8 所示。即使他们已经由于其他的理由而停止，此时审查客户也可能帮助理解不同类型的停止。

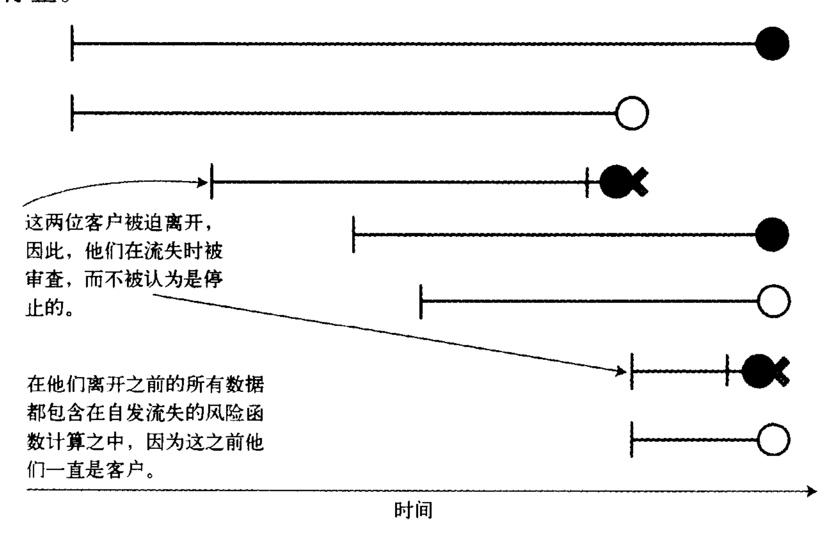


图 12-8 使用审查，使开发包含强制离开的客户的自发流失风险模型成为可能

12.3 从风险到生存

本章从保持曲线的讨论开始。利用风险函数，可以产生一条非常相似的曲线，称为生存曲线。生存曲线更有用，并且在很多意义上感觉更精确。

12.3.1 保持

保持曲线提供了关于在某一段时间内保持了多少客户的信息。产生保持曲线的一个通常的方法如下：

- 对 1 周之前开始的客户，测量 1 周的保持；
- 对 2 周之前开始的客户，测量 2 周的保持；
- 依此类推。

图 12-9 展示了基于这种方法的一条保持曲线示例。该曲线的总体形状看起来是恰当的。然而，曲线本身参差不齐，比如，看起来很奇怪的是，数据显示，10 周的保持会比 9 周的保持更好。

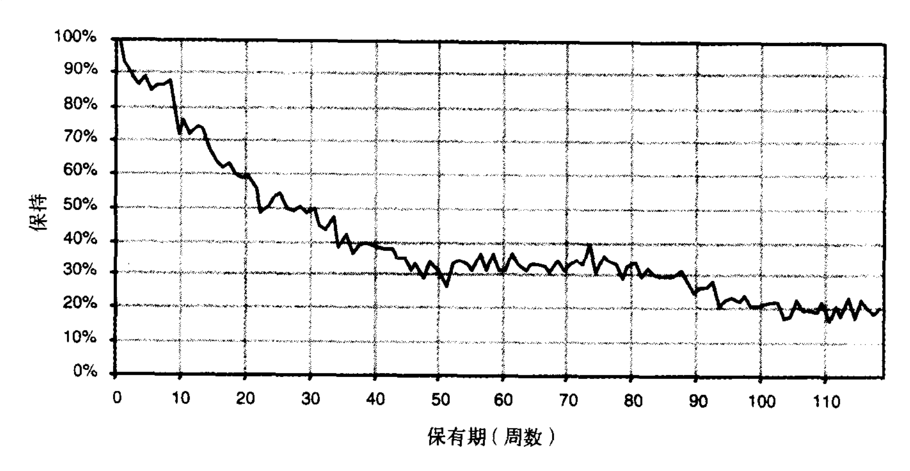


图 12-9　一条保持曲线可能参差不齐

事实上，不止是形状奇怪，它违犯了保持的最主要的观念。例如，它开启了曲线将会多次与 50%的阈值相交的可能性，导致怪异的、不准确的结论——有不止一个中值生存期，或者是，在客户关系开始后，前 10 周的平均客户保持可能超过前 9 周的平均值。这是怎么回事？难道客户转世重生了？

这些问题是在产生曲线的过程中人为造成的。任何给定时间段获取的客户可能比其他时间段获取的客户更好，也可能更差。例如，也许 9 周前，有一个特殊优惠价格服务引入了差的客户，而在 10 周以前开始的客户是好的和差的混合体，但是那些 9 周以前开始的客户特别差。因此，9 周以后差的客户少于 10 周后较好的客户。

客户的质量也可能仅仅由于随机变化而发生改变。毕竟，在前面的图中，考虑的是 100 多个时间段——因此所有的事情都是平等的，有些时间段预期会有所差异。

一个复杂的原因是，市场营销工作随着时间而变化，从而吸引不同质量的客户。例如，来自不同渠道的客户通常有不同的保持特征，并且来自不同渠道的客户混合体可能随时间而改变。

12.3.2 生存

风险给出了客户可能在某个时间点停止的概率。另一方面，生存提供客户保持到该时间点的概率，生存值可以直接从风险来计算。

在任何时间点，客户生存到下一个时间单元的机会简单地说就是1-风险，称为时刻 t 的条件生存（它是有条件的，因为它假设客户生存到时刻 t）。计算给定时间的全部生存需要累积到该时间点的所有条件生存，把它们相乘。生存值在时刻 0 的初值为 1（或 100%），因为在分析中所有客户在分析的开始都生存。

因为风险总是介于 0 和 1 之间，所以条件生存也是处于 0 和 1 之间。因此，生存本身总是在变小——因为每个相继的值都乘以一个小于 1 的数。生存曲线从 1 开始，逐渐地下降，有时可能变平，也可能消失，但是从不上升。

对于客户保持目的来说，生存曲线比前面描述的保持曲线更有意义。图 12-10 展示一条生存曲线和相应的保持曲线。显然生存曲线是平滑的，而且一直下降，而保持曲线在所有位置上下跳跃。

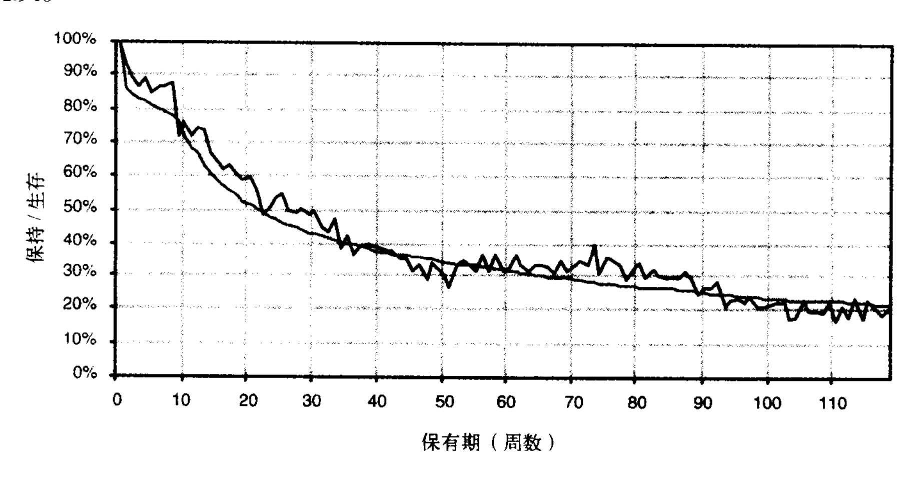


图 12-10 生存曲线比保持曲线更平滑

初看起来，保持曲线和生存曲线的差别可能不是直观的。保持曲线实际上是把从过去开始的一整串客户的不同图像粘贴到一起，就像由一串不同的照片拼凑成的抽象拼贴画而得到的一幅全景图。在抽象拼贴画中，每张照片中的图像都相当清晰，然而边界未必平滑地过渡。抽象拼贴画的不同照片看起来是不同的，这种不同是因为光线不同或者归因于抽象拼贴画美学的视图差异。

同样的事情在保持曲线中也会遇到，在保持曲线中，开始于不同时间点的客户有不同的视角。保持曲线上的任何给定点都接近真实的保持值；然而，作为一个整体，它看起来参差不齐。一种除去参差不齐的方法是关注同时开始的客户，就像本章前面建议的那样。然而，这大大减少了作用于曲线的数据量。

提示： 不要使用保持曲线，使用生存曲线。即首先计算风险，然后回来计算生存曲线。

另一方面，生存曲线可以考察尽可能多的客户，而不仅是恰好 n 个时间段之前开始的

那些客户。任何给定时刻 t 的生存就利用了来自所有客户的信息。在时刻 t 的风险使用保有期大于等于该值的所有客户的信息（假设所有人都在风险人口中）；然而，生存是从 t 的较小值开始，结合风险的所有信息来计算的。

因为生存计算使用所有的数据，所以得到的值比保持计算更稳定。保持曲线的每个点把客户限制在开始于某个特定时间。另外，由于生存曲线总是呈下降趋势，所以客户半衰期和平均客户保有期的计算更精确。通过合并较多的数据，生存可以提供关于客户保持的更精确平滑的图像。

当分析客户时，风险和生存都提供了关于客户的有价值的信息。因为生存是累积的，它给出了比较不同群体客户的好的概要数值：对于不同的群体，1 年的生存比较情况如何？生存也用作计算客户的半衰期和均值客户保有期，从而反过来馈入其他的计算，如客户价值等。

生存是累积的，很难看到特定时间点的模式。而风险使特殊的原因变得更加明显。当讨论一些现实世界的风险时，有可能识别出客户生存周期中增加风险的事件，生存曲线对这些事件的突显不像风险那么清楚。

在比较不同客户群体的风险时也可能发现，比较一段时间的平均风险没有意义。从数学的角度看，“平均风险”没有意义，正确的方法是将风险转变成生存，在生存曲线上比较那些值。

迄今为止给出的风险和生存的描述与统计学中对这一问题的处理略有不同。下面“关于生存分析和统计学的注解”部分更进一步解释这种差异。

关于生存分析和统计学的注解

本章中关于生存分析的讨论假设时间是不连续的。尤其，事件发生在某些特别的日期，而那一天是哪个特别时间并不重要。这不仅对于数据挖掘所提出的问题是合理的，它看起来也更直观，而且也简化了数学问题。

但在统计学中，生存分析做出的是相反的假定：时间是连续的。统计学家使用的不是风险概率，而是风险比率，使用指数和加和把它转化为生存曲线。在比率和概率之间的一种区别是比率值会大于 1，然而概率永远不会。同时，比率对碰到的许多客户生存问题看上去不那么直观。

本章计算风险的方法称为寿命表格法，且它对于不连续的时间数据工作良好，另一个非常相似的方法称为 Kaplan-Meier 法，常用于连续时间数据，当事件的发生时间不连续时，这两种技术几乎产生同样的结果。

统计生存分析的一个重要部分是利用参数化回归方法进行风险估计——试图从风险中找到最好的函数形式，这是另外一种可选择的方法，它可以从数据中直接计算风险。

这种参数化方法的重要优势是它能够更容易地把共同变量包括到处理过程中。本章稍后有一个基于这种参数化模型的例子。不幸的是，风险函数很少遵循非统计学家熟悉的形式。风险对于描述客户寿命周期非常好，因而如果一个简单的函数可以捕捉到如此丰富的复杂事物，那将会令人非常吃惊。

我们强烈鼓励有数学或统计学背景的读者在该领域进行更深入的研究。

12.4 比例风险

David Cox 爵士是 20 世纪最权威的统计学家之一，他的著作包括许多书籍和 250 多篇论

文。他获得过许多奖项，包括 1985 年伊丽莎白女王颁发的爵士头衔。他的很多研究内容是以理解风险函数为中心，而且他的工作在世界医学研究领域有着特殊的地位。

他的开创性论文是关于在风险方面确定最初因素（零时共同变量）的影响。假定这些最初的因素在风险方面有一个统一的比例效应，他能够找出如何测量不同因素的影响，本节的目的是介绍比例风险，理解它们对于理解客户非常有用。本节首先用实例说明为什么比例风险有用，然后介绍另一个替代的方法，最后回到 Cox 模型本身的讨论。

12.4.1 比例风险实例

考虑下列一个关于吸烟危险的陈述：*吸烟者得白血病的风险比不吸烟者高 1.53 倍*，该结果是一个关于比例风险的著名实例。在研究这个问题的时候，研究者已经知道某人是否是吸烟者（事实上，还存在第三组，即从前吸烟者，但我们这里的目的是举例说明一个实例）。某人是否是吸烟者是一个初始条件的例子，因为只有两个可考虑的因素，所以可以仅仅看一下风险曲线，就可以得到总风险的某种平均值。

图 12-11 提供了一个来自市场销售界的图示，它显示了两个风险概率集合，一个是受电话诱惑加入的客户，另一个是通过直接邮寄加入的客户。再次强调，某个人如何成为一位客户是一个初始条件的例子。电话推销客户的风险是非常高的，观察一下图表可以发现，电话推销客户的风险比直接邮寄客户几乎高出两倍。Cox 比例风险回归提供了量化这个问题的方法。

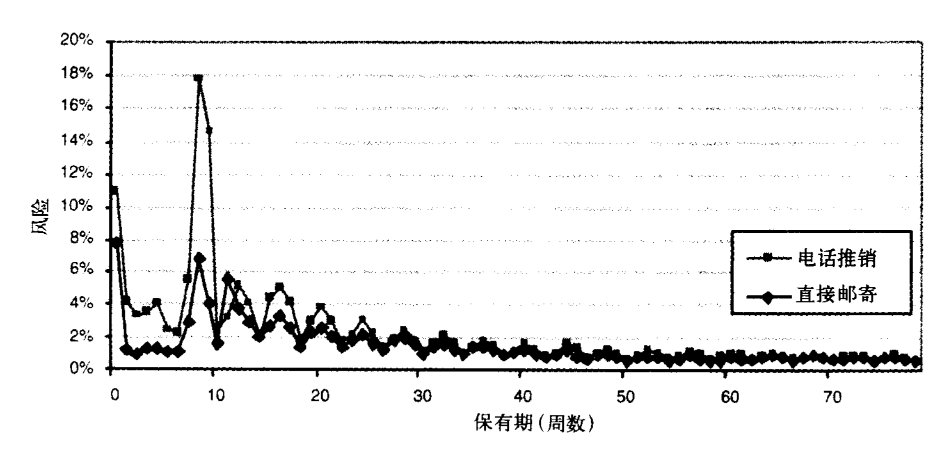


图 12-11 这两个风险函数显示，通过电话推销获得的客户流失风险大约比直接邮寄高 1.5 倍

刚刚提到的两个例子使用分类变量作为风险因子。我们可以考虑另外一个关于香烟风险的陈述：*每年每吸食一包香烟，得结肠癌的风险率就增加 6.7%*。这一陈述与前例不同，因为它现在取决于一个连续变量。应用比例风险，可能确定分类变量及连续变量二者的贡献大小。

12.4.2 分层：测量生存的初始结果

图 12-11 显示了两个不同客户群组的风险概率，一个是电话推销活动，另一个是直接邮寄营销活动，这两条曲线清楚地显示了两种渠道之间的区别。利用 1 年生存期、中值生存期或平均截短保有期，可以为这些风险生成一条生存曲线，并量化这种差别。这种测量按照初始条件定义的不同组之间差别的方法，被称为*分层*，因为每一组的分析均是独立于其他组进

行的，这样就产生好的可视化且精确的生存值。这当然也很容易，因为一些统计包，如 SAS 和 SPSS，有这样的选项，这些选项使得为这一目的给数据分层变得很容易。

分层解决了假定两种条件为真时，理解初始结果的问题。首先，初始结果必须是一个分类变量，因为这些数据必须被分解为独立的群组，某些变量如渠道、产品或区域等，需要被选择用于这一目的。当然，使用归档将连续变量分解为不连续块也总是可以的。

第二，每个群组必须相当大，当开始使用许多客户且仅用一个取少数几个值的变量（如渠道）时，这不成问题。然而，可能存在令人感兴趣的多重变量，如：

- 获取渠道
- 初始提升
- 地理状况

一旦包括一个以上的维，分类的数目会增加很快，这意味着数据会逐渐稀疏地展开，使风险估算变得越来越不可靠。

12.4.3 Cox 比例风险

1972 年，David Cox 爵士发现了这个问题，并提出一个分析方法，现在我们称之为 Cox 比例风险回归法（Cox proportional hazards regression），这种方法克服了这些局限性。他杰出的洞察力就是找到了一种方法以关注初始条件而不是风险本身，这个问题是：初始条件对于风险会有什么样的影响？他解决这个问题的方法非常耐人寻味。

幸运的是，这一想法比他解决该问题的数学方法更简单，他关注的不是风险，而是引入了局部可能性的思想。假定在给定的时间 t 内只有一位客户停止，那么在时间 t 内的局部可能性就正好是那个特定客户停止的可能性。

对局部可能性的计算是用代表某个特定客户停止风险的任何函数或数值除以该时间内可能停止的所有客户风险总和。如果所有客户具有相同的风险比率，那么，这个比率将会是一个常数（1 除以那个时间点的总人口）。然而，风险不是常数，但愿是某些初始条件的函数。

Cox 做的一个假设是初始条件对于所有风险有一个不变的影响，不考虑风险随时间的变化。局部可能性是一个比率，比例性假定的意思是，无论风险是什么，风险都会同时出现在分子和分母中，基于初始条件乘以一个复杂的表达式，结果就是一个包含初始条件的复杂数学公式。风险本身已经从局部可能性消失，它们彼此完全抵消。

下一步是应用所有停止客户的局部可能性来得到这些特定客户停止的总体可能性，所有的这些局部可能性的乘积表示：当客户确实停止时，准确观察到停止客户停止过程中一个特定集合的可能性。方便的是，这种可能性也可以仅仅表示成初始条件的函数，而不是风险的函数，风险可能是未知的。

幸运的是，有一个称为极大似然估计的统计学领域，即当给出一个类似事件的复杂表达式时，它可以找到参数值，使得结果成为最大可能。这些参数值可方便地表现出这些初始值对于风险的影响。作为一个额外的奖励，这种技术可同时用于连续数值及分类数值，而分层法仅适用于分类数值。

12.4.4 比例风险的局限性

Cox 比例风险回归法是非常有力且非常智慧的方法，但也有其局限性。为了让该方法工

作良好，Cox不得不做出许多假设。他围绕连续时间风险设计他的方法，而且假设在任何给定的时间内只有一位客户停止。通过某些调整，比例风险回归的实施通常对不连续时间风险有用，而且可以在同一时间内处理多重停止。

警告：Cox比例风险回归把初始条件对于整个风险函数的影响分级和量化。然而，该结果高度依赖于通常可疑的假设，即初始条件对于风险在整个时间内都有不变的影响。所以使用它要小心。

在比例风险模型中最大的假设是对比例本身的假设，即，初始条件对于风险的影响不具有时间成份。实际上，完全不是这样的。即使曾经有过，初始条件也很少会有如此完美的比例，即便是在科学领域。在市场营销领域，可能性甚至更小，市场营销不是一个可控制的实验，情况一直在发生变化，新的计划、定价变化及竞争时常发生。

坏消息是，考虑到整个过程的不同影响，没有一个简单算法可以解释初始条件；好消息是，这通常没有什么差别。即使利用比例假设，Cox回归在以下方面仍工作良好，即决定哪些共同变量对于风险有一个大的影响，换句话说，它在解释什么样的初始条件与客户离开是相互关联方面工作良好。

Cox的方法只为零时共同变量设计的，就像统计学家所说的初始值。这种方法已经扩展为处理在客户生存期发生的事件，比如他们是否升级产品或有所抱怨。用统计学的术语来说，这些都是依赖时间的共同变量，是指附加因素在客户保有期内的任一点都可能发生，而不仅在关系的初期。这样的因素可能是客户对保持活动的响应或客户的抱怨。由于Cox的开创性工作，他和其他一些统计学家已经拓展这项技术，使其包括这些类型的因素。

12.5 生存分析实践

从客户保持的角度来看，对于了解客户及量化市场营销工作来说，生存分析已经被证明是非常有用的，它提供一种方法来估算在某些事情发生前它将保持多长时间。本节将给出一些生存分析的特殊实例。

12.5.1 处理不同的流失类型

与客户打交道的公司必然会涉及到由于各种原因造成的客户离去，在本章前面的部分，已经描述了风险概率，解释风险如何阐明那些影响客户生存周期的企业各个方面。特别是，风险峰值与强制那些没有付清账单的客户尽快离去的商业过程是相一致的。

由于这些客户需要不同地对待，将他们从风险计算中完全移除的尝试是错误的方法，问题在于，只有在客户已经被迫停止之后，才会知道要移去哪些客户。就像前面提到的，应用在客户关系结束时获得的知识来过滤即将分析的客户，不是一个好主意。

正确的方法是把这个问题分解为两个问题：自发流失的风险是什么？强制流失的风险是什么？其中每一个问题都使用所有的客户，审查由于其他因素离开的客户。当计算自发流失的风险时，无论客户何时被强迫离开，该客户仍包括在分析过程中直到他或她离开为止——在那一点，该客户被审查。这是有道理的，因为一直到客户被迫离开这一刻，该客户都没有主动离开。

这一方法可以进行拓展以便用于其他目的。以前，本书的作者试图了解一家报纸的不同客户组，特别是，按照获取渠道进行的生存分析如何按时间改变或者不变。不幸的是，在一

个时期内，发生了一次联合抵制这份报纸的活动，在那段时间总体停止水平提高了。毫无疑问，这段时间风险上升，生存降低。

有方法来考虑这些特殊的停止吗？回答是肯定的，因为公司很好地记录了客户停止的原因。那些联合抵制该报的客户在停止的当天被简单地进行了审查——正如在医学界所说的，这些客户已经没有必要追查到底。通过审查，可能得到一个在没有联合抵制的情况下对总体风险的准确估计。

12.5.2 客户何时会回来

迄今为止，对于生存分析的讨论一直聚焦在客户关系的结束。除了预测坏事情发生的概率之外，生存分析还可用于许多事情，例如，可用于估计客户停止后何时会返回。

图 12-12 显示了一条生存曲线，展示了客户在停止使用移动电话服务之后再次启用的风险。在这个实例中，风险是一个给定停用天数后客户返回的概率。

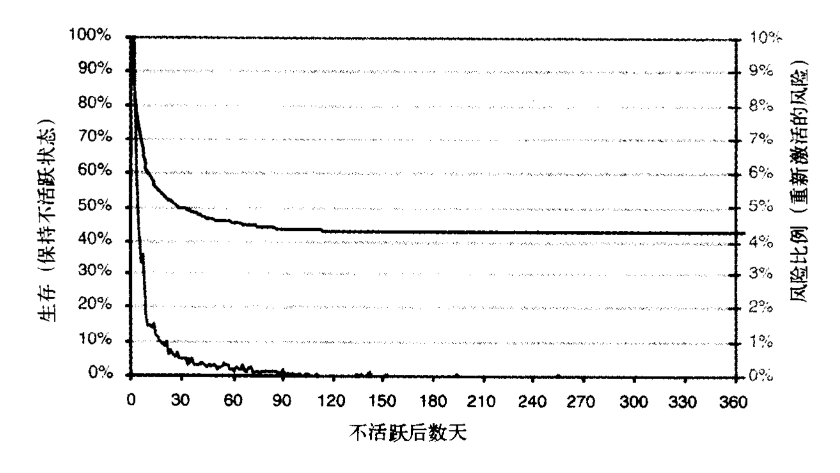


图 12-12 移动电话客户重新激活的生存曲线（较高的曲线）和风险曲线（较低的曲线）

这些曲线有几个重要的特征。首先，最初的重新激活比率是非常高的。在第一周内，超过三分之一的客户重新激活。商业规则可以解释这一现象，许多停机是因为客户没有支付账单，这些客户中的大部分仅仅是坚持到最后一刻——他们实际上想保留电话，只是不喜欢支付账单。然而一旦电话被停机，他们便很快付清全部账单。

90 天以后，风险实际上等于 0——客户不能再激活。商业过程又一次提供了指导：电话号码在客户离开以后会保留 90 天。通常地，当客户重新激活使用，他们希望保留同一个电话号码，而 90 天以后，号码可能已被重新分配，客户可能会得到一个新号码。

这项讨论掩饰了一个问题，那就是，新的（重新激活）客户是如何与过期账户相关联的。在这个实例中，分析过程把电话号码与一个账户 ID 关联使用。这非常好地保证了匹配的准确性，因为重新激活客户保留了他们的电话号码及账单信息。这有些保守，但对于找到重新激活的情况是起作用的。它对于找到其他类型的赢回情况不起作用，比如那些为了得到初期折扣而愿意更换电话号码的客户。

另一种方法是试图识别不同个体随时间的变化情况，即使他们属于不同的账户。对于那些把收集社会安全号（Social Security Number，SSN）或驾驶执照号码作为其业务领域的常

规部分的商务来说，这样识别的号码能够与账户随时间的变化情况连接起来（要知道，并不是每个被要求提供这种识别信息的人都做得那么准确)。有时候，匹配姓名、住址、电话号码及（或）信用卡对于匹配目的就足够了。但更通常的情况是，这项任务被转包给一家分配个人和家庭 ID 的公司，然后提供需要的识别信息，找出哪些新客户是已经被赢回的真正的前客户。

研究初始共同变量增加了更多的信息，在这种情况下，“初始”的意思是关于客户停止活动所在点的任何已知信息。这不仅包括像初始产品和促销信息，也包括客户在停止活动之前的行为。牢骚满腹的客户是更可能还是更不可能重新激活？漫游的客户怎样呢？迟付账单的客户呢？

这一实例显示了利用风险了解一个经典的“时间事件”问题。生存分析能够处理的其他一些此类问题是：

- 如果客户开始于一个最低费率计划，在他们升级到高级费率计划之前要多长时间？
- 客户何时会升级到一个高级费率计划，在他们降级之前会有多长时间？
- 已知过去的客户行为和不同的客户有不同的购买周期的事实，客户购买时间间隔预期有多长？

利用生存分析的一个好处是，可以很容易地查询不同初始条件的影响——比如在过去的时间里一位客户访问过的次数。使用比例风险，可以确定哪些共同变量对于预期的结果最有影响，包括哪种干预是最可能还是最不可能起作用。

12.5.3 预测

生存分析的另一个重要的应用是预测未来客户的数目，或者说，在未来给定的一天中停止的客户数目。总体来说，对于估计在给定的时间长度内有多少客户将会保留，生存分析是很有效的。

对于任何这样的预测都有两个组成要素，第一是一个当前客户模型，它可以考虑到客户生存周期期间多种共同变量。这样的模型通过把一个或者多个生存模型应用到所有客户而工作。如果一位客户已经存在了 100 天，那么明天停止的概率就是第 100 天的风险。为了计算后天停止的可能性，首先假定客户在明天不停止，然后在第 101 天确实停止，则后天停止的可能性就是第 100 天的条件生存率（1 减去风险——不停止概率）乘以第 101 天的风险。将这个概念应用到所有客户保有期，就可能预测现有客户未来的停止情况。

图 12-13 显示了这样一个对 1 个月内停止的预测，它是由生存专家 Will Potts 开发的，同时给出的还有在这段时期观察到的真实值，以生存为基础的预测被证明与实际发生的事情相当接近。顺便说一句，这种特殊生存估计使用一个风险参数模型，而不是经验的风险估计，该模型能够考虑到每周的不同工作日。周循环中停止情况的结果在图中可以很明显看出。

客户层次的预测的第二个组成要素要计算起来有一点困难，这个要素就是新客户对预测的影响，但困难不是来自技术上的。我们所面临的挑战是对新的起点进行估计。幸运的是，通常的预算预测包含新的起点，有时按产品、渠道或者地理状况等分解。把这些影响考虑进来对生存模型进行精修是可能的，当然，这种预测的准确性只能与预算的准确性一样高。尽管最理想的情况是，这种基于生存分析技术的预测能够融入到根据预算水平管理实际水平的过程。

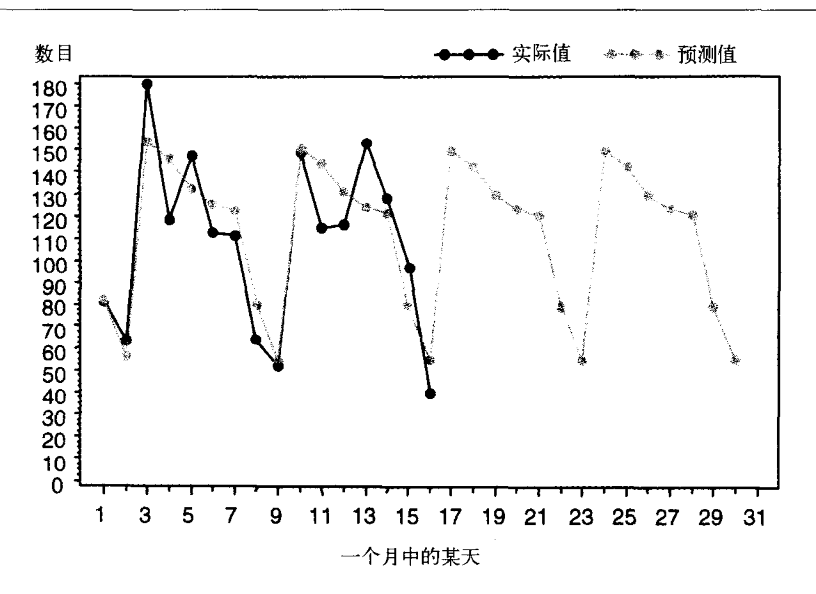


图 12-13 生存分析也能够应用于预测客户停止

这些要素的结合——对现存客户停止的预测及对新客户停止的预测——使得对未来客户层次的估计成为可能。作者一直在与那些向前预测数年的客户一起工作。因为这些针对新客户的模型包括了获取渠道，所以预测模型可能用于优化未来的多种获取渠道组合。

12.5.4 风险随时间变化

在生存分析中一个更加困难的问题是，风险本身是否是持续不变的，或者说它们是否随时间而变化。在科学研究中假定风险不发生变化，科学研究中生存分析的目的是获得在不同情形下对“真实”风险的估计。

这种假设在市场营销工作中可能真也可能不真。当然，通过这些假设，生存分析利用客户数据已证明了它的价值。然而，考虑风险随时间而变化的可能性也是很有意义的。特别是，如果风险确实发生了变化，那么它就会给出某些启示，告诉我们营销地点和客户随时间的变化情况到底是在变好还是变坏。

要回答这个问题，一个方法是使风险基于停止的客户而不是开始的客户，特别是，比方说，那些在过去的几年内每年停止的客户。换句话说，把去年停止的客户相关联的风险与前年停止的客户相关联的风险相比较，是否有显著的不同？在本章的前面部分已经提醒大家，对于一个按照停止数据选定的客户集合计算风险不会得到准确的风险。应该如何克服这个问题呢？

有一种计算这些风险的方法，虽然这还没有在标准的统计工具中出现过。这种方法对客户使用时间窗来估计风险概率。让我们回忆一下经验风险概率的定义：在某个特定时间实际停止的客户数除以在那个时间可能停止的客户数。到目前为止，所有的客户都被包含在计算之中。这种方法的目的是只把客户限制在那些在研究期间可能会停止的客户。

作为一个实例，我们基于 2003 年停止的客户来估算风险。在 2003 年停止的客户要么是在 2003 年第一天是活跃的客户，要么是那年的新客户。无论哪种情况，这些客户都对人口总数做出了贡献，无论他们的保有期是否从 2003 年第一天算起（对于新客户是 0）。

让我们考虑 1 天的风险概率计算。那些保有期为 1 天、可能停止并且在 2003 年确实停止的客户数到底是多少？只有那些在 2002 年 12 月 31 日到 2003 年 12 月 30 日之间开始的客户有可能在 2003 年有一个 1 天的停止。因此，一天的风险计算使用在 2003 年保有期为 1 天的所有停止作为停止总数，风险人口由 2002 年 12 月 31 日到 2003 年 12 月 30 日之间开始的客户组成。作为另外一个例子，365 天的风险可能会以在 2002 年开始的客户人口总数为基础。

得到的结果就是以某个特定时段的停止为基础的风险估计。从对比的角度来看，生存被证明比风险本身更有用。图 12-14 给出了一个例子，表明生存在那几年的过程中的确在下降。生存方面的改变很小，但计算是以数十万计的客户为基础，确实表明客户质量的下降。

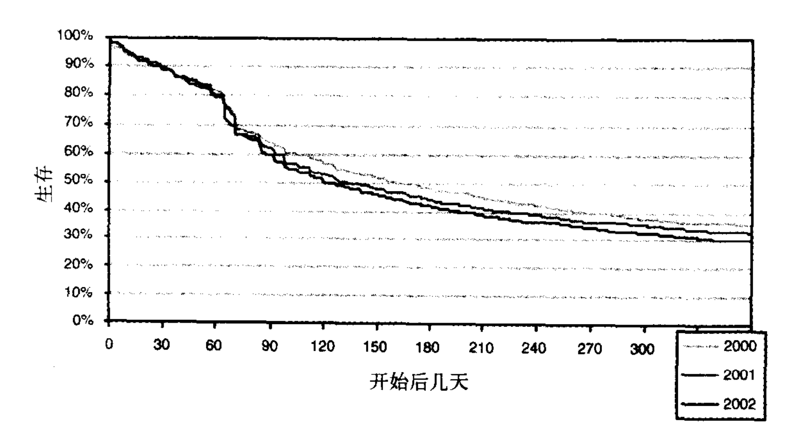


图 12-14 时间窗技术使得观察生存随时间的变化成为可能

12.6 小结

风险和生存分析是为了解客户而设计的，本章引入了风险作为客户在某个给定时间点离开的条件概率。这种对于生存分析的处理按照统计学是不正统的，在统计学中更喜欢基于连续的比率而不是离散的时间概率，但这种处理方法对于分析客户更直观。

风险对于客户生存周期起到了像 X 光片的作用（即可以提早发现潜在的风险）。这个与生存相关的概念（即到某个时间点仍然存活下来的客户比例），使得比较不同客户组以及把这些结果转换成经济效益成为可能。当有足够多客户的时候（通常是有的），通过为每一个客户组建立一条曲线来把客户分层，提供了一种好的比较方法。利用其他度量，比如，某个特定时间点的生存、客户半衰期、平均保持时间，也可能更好地了解客户。

生存分析中关键的概念之一是审查过程，这就是说某些客户在分析过程中会离去。这一审查观点可以被拓展用于对竞争风险的理解，如自发流失和强制流失的问题。审查也可能用于舍弃某些结果（比如一次联合抵制活动），不至于对总体结果造成有害的偏离。

风险的最有力方面之一是，在过程的开始，确定哪些因素对于风险的增加或减少是可信的。除了客户分层技术之外，还有另外的技术——Cox 比例风险回归，20 世纪 70 年代以来它已经证明了自身的价值，且不断扩展并得到改进。

除了测量客户离开的概率以外，生存分析还有更多的应用。它已经被用于预测客户层次以及客户生存期中其他类型的事件。它是一个强有力的工具，似乎是专门为了解客户及他们的生存周期而设计的。

第13章 遗传算法

就像基于存储的推理和神经网络一样，遗传算法（genetic algorithm，GA）也是以模仿生物过程为基础。在数百万年之间，进化和自然选择已经造就了对环境高度适应的特殊物种。通过将一代最适应的生物个体中的遗传物质传播给下一代，进化论能够优化下一代个体的适应度（fitness）。

遗传算法将相同的观念应用到一些其结果能表示为最优“个体”，目标是最大化个体的“适应度”的问题。许多问题都可以这样描述，挑战在于用合适的方式编码这些问题。举例来说，遗传算法的应用之一是训练神经网络。此时个体就是网络内部的一组权重值，个体的适应度是具有训练集上的那些权重的神经网络的准确度。训练以进化方式进行，使更适应的个体将权重传播到下一代。不太适应的个体及其遗传物质不再保存。尽管偶然性在任何特定个体的生存中起着非常重要的作用，但在一个较大的群体中，对自然选择来说有足够多的不同类型的个体例子，传播能够产生最适应个体的遗传物质。

遗传算法也称为进化算法，已经被应用到各种不同行业的优化问题，包括复杂的计划安排问题、大型工厂的资源优化问题和包括复杂数据类型的分类问题。也与其他的数据挖掘算法结合使用，包括用来确定神经网络的最佳拓扑，确定基于存储的推理的得分函数，以及前面提到的优化神经网络的权重。然而，在一般的数据挖掘软件包中普遍没有遗传算法。

优　化

优化问题（optimization problem）有三个特征：

- 一组参数（遗传算法称为基因组或染色体）。
- 一个函数（适应度函数），把多个参数组合成一个单一的数值。
- 在参数上的一系列限制（对于遗传算法，这些已经并入适应度函数）。

目标是寻找使适应度函数最大或最小的参数，并服从限制。即使对最先进的计算机来说，搜遍所有符合限制的参数组合也是很麻烦的；即使对于少数的几个参数，组合后的数目仍然还是太大而无法搜寻。

遗传算法是解决这类问题的一种方法，但不是惟一的方法。当适应度函数满足一些特殊的数学条件时，微分学能用来寻找最优解。尽管在实践中极少函数是可微分的，但是微积分学也包含在其他情况下估计解的思想。用于训练神经网络的共轭梯度方法（conjugate-gradient method）就是基于这样的思想。就像“Excel 的计算器功能”。

另一种方法发生在线性规划问题中。在这些问题中，适应度函数是线性的，而且所有的限制也是线性的。这些限制时常出现在资源分配问题中，诸如：

公司在一组工厂生产小装置，每个工厂有生产量、产品成本和运送小装置到客户的花费。每个工厂应该生产多少小装置才能以最低成本满足客户的需求？

解决这类问题的标准方法称为单形法（simplex method），而且它在计算方面是可行的。这类问题已经用数以千计的变量解决了。线性规划类型问题的更多信息见网站 www-unix.mcs.anl.gov /otc /Guide /faq /linear-programming-faq.html。

另一种方法称为模拟退火（simulated annealing），即模拟物理过程：液体冷却并形成晶

体的模式。晶体最小化特定类型的能量，而且贯穿整个结晶过程。研究物理性质的科学家最经常使用模拟退火方法。

遗传算法的第一项工作始于20世纪50年代后期，当时生物学家和计算机科学家合作，为早期的计算机上的进化机制建立模型。稍后，在20世纪60年代早期，密歇根大学的John Holland教授和他的同事们把计算遗传学方面的工作，包括染色体、基因、等位基因和适应度函数等，应用到优化问题。1967年，Holland的一位学生J.D.Bagley在其毕业论文中首次提出了用遗传算法来描述优化技术。当时，由于遗传算法在进化解的过程中依赖于随机的选择，许多研究员对遗传算法感到不舒服；这些选择似乎是随意的和不可预知的。在20世纪70年代，Holland教授发展了该项技术的理论基础。他的模式（schema）理论提供了为什么使用遗传算法的深入了解，而且让人感兴趣的是，他提出了遗传学本身为什么能够创造像我们自己一样的、成功的并且能适应的创造物。在数据挖掘和数据分析界，遗传算法不像其他技术一样使用广泛。数据挖掘关注像分类和预测之类的任务，而不是优化。尽管许多数据挖掘问题能够设计为优化问题，但是这不是平常的描述。举例来说，一个典型的数据挖掘问题可能是以第一个星期的销售为基础，预测一个目录中给定项需要的存货层次、目录中项的特征和容器。把它改述为一个优化问题，就变成有几分像“对预言性目的来说，什么函数最适应存货曲线”。应用统计学回归技术（statistical regression technique）是寻找该函数的一种方法，将数据回馈到一个神经网络是另一种估计的方法，使用遗传算法也提供了一种方法。前面“最优化”部分讨论了为这一目的而特别设计的其他一些方法。

本章包含计算机上的遗传学的背景，并且介绍了由John Holland设计的模式机制，解释为什么遗传算法起作用。本章主要讨论两个案例研究，一个是在资源优化（resource optimization）领域，另一个是在预测邮件消息方面。尽管目前只有少数商业数据挖掘产品包含遗传算法，但是更多的特殊软件包确实支持该算法。它们是一个重要的、活跃的研究领域，而且未来可能会得到更广泛的应用。

13.1 遗传算法如何工作

遗传算法的能力来自其生物基础，进化论已经证明，适者生存（见后面“遗传学的简单概观”部分）。成功地绘制人类基因组的模板，即被人类个体共享的所有常见DNA，仅仅是开始。人类的基因组已经在许多领域，像医学研究、生物化学、遗传学，甚至人类学中提供先进的知识。人类基因组虽然很重要，已经超出需要理解遗传算法的知识范围，但是描述计算机技术需要的语言过去一直大量地借鉴生物模型，如下所述。

13.1.1 计算机上的遗传学

一个简单的例子有助于说明遗传算法如何工作：设法找出有单一整数参数 p 的简单函数的最大值。本例中的函数是由 $31p-p^2$ 定义的抛物线（看起来像颠倒的“U”），其中 p 的变化范围在0到31之间（见图13-1）。参数 p 被表示成一个含有5个二进制位的字符串，代表从0到31的数字；这个位串就是遗传物质，称为基因组。适应度函数在15和16的峰值，分别表示为01111和10000。这个例子说明，即使有多重不同的峰值，遗传算法仍然是适用的。

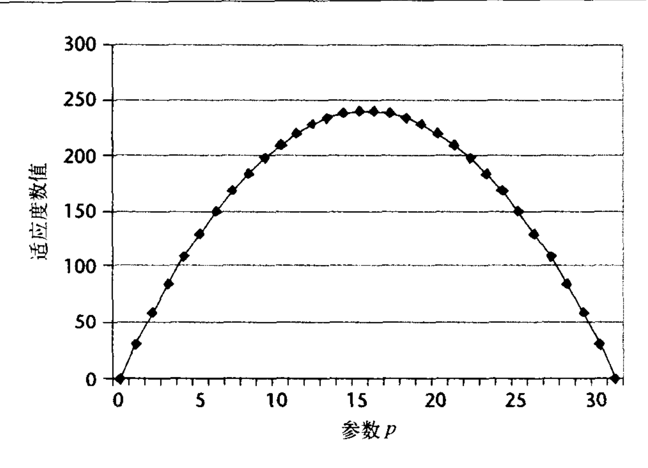


图 13-1 寻找这个简单函数的最大值有助于阐明遗传算法

遗传算法通过进化基因组，来得到越来越适应的后代；也即，提供更好的解决问题的办法。在自然界中，适应度只是生物体的生存和繁殖能力。在一台计算机上，进化模拟下列几步：

1）识别基因组和适应度函数；

2）产生一个初始基因组的代；

3）通过应用遗传算法修改初始代；

4）一直重复第 3 步，直到群体的适应度不再改变。

第一步是表达问题。在这个简单的例子中，基因组由参数 p 对应的一个单独的、5 个二进制位的基因组成，适应度函数是抛物线。在代与代之间，适应度函数将被最大化。

对于这个例子，如表 13-1 所示，初始代包含四个随机产生的基因组。通常，一个真正的待处理群体会有数百或数以千计的基因组，但是对这里的说明目的是不实用的。注意，在这个群体中，平均适应度是 122.5，已经相当好了，因为实际的最大值是 240，但是进化能改进它。

表 13-1 四个随机生成的基因组

基 因 组	p	适 应 度
10110	22	198
00011	3	84
00010	2	58
11001	25	150

基本算法使用三个操作修改初始群体：选择（selection）、交叉（crossover）、变异（mutation），如图 13-2 所示。这些操作在下面解释。

遗传学的简单概观

生命依赖于蛋白质，蛋白质由 20 个称为氨基酸的基本单元的序列构成。细胞核的染色体是携带细胞需要的蛋白质的蓝图 DNA 序列。每个人的细胞的 23 对染色体一起构成这个人的基因组。大体上，同一物种不同个体的基因组彼此很相似，然而，确实有个体的差异。

基因组中的 DNA 使用核苷酸序列编码这些氨基酸序列蓝图。这些核苷酸构成遗传基因字母表的四个字母：

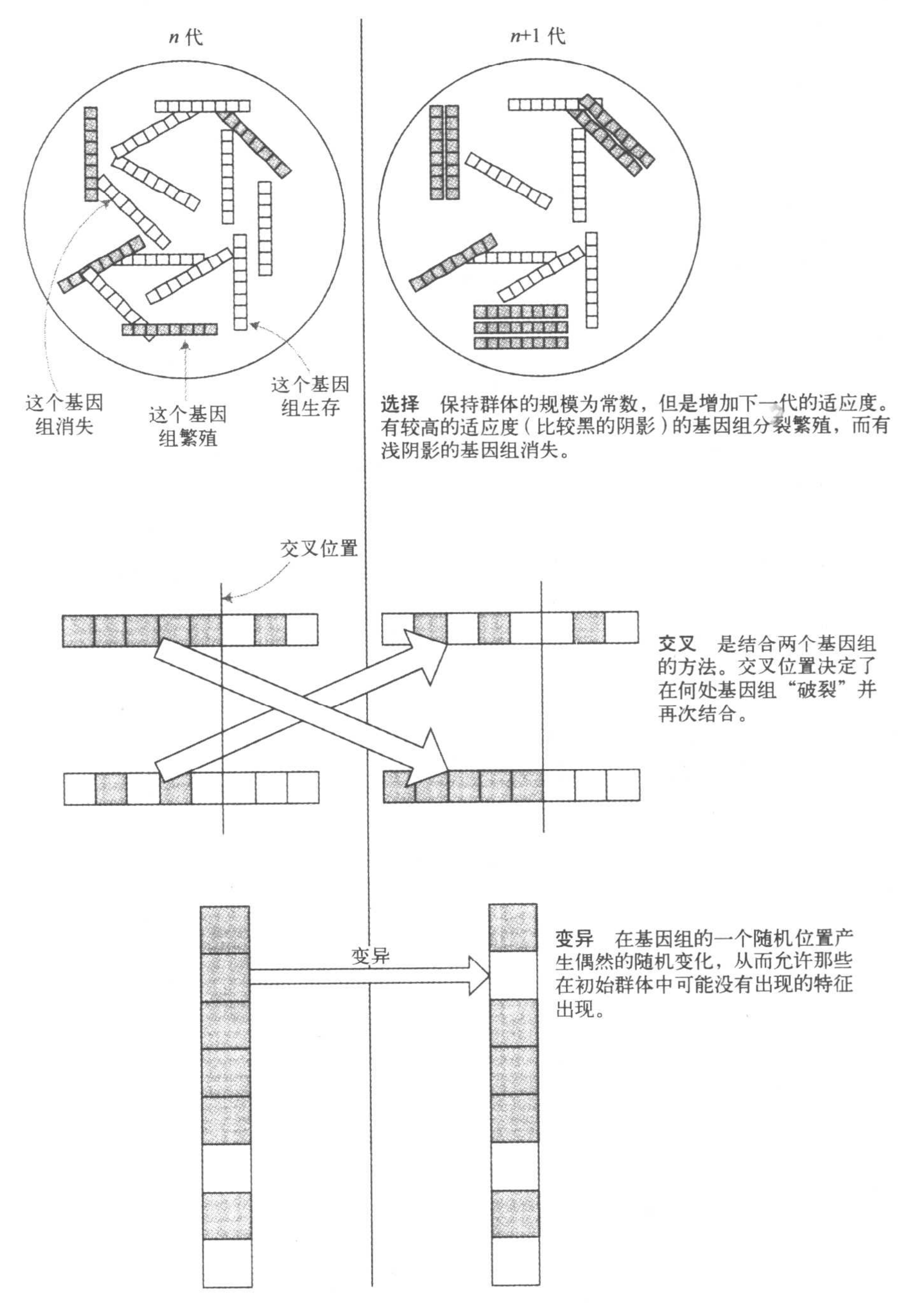


图 13-2 遗传算法的基本操作是选择、交叉和变异

- A，腺呤
- C，胞嘧啶
- G，鸟嘌呤
- T，胸腺嘧啶

核苷酸用三元组表示20个氨基酸。举例来说，被称为甲硫胺酸的氨基酸对应三元组ATG。另外的一个氨基酸——赖胺酸，有两种拼法：AAA和AAG。因此，如果一个DNA

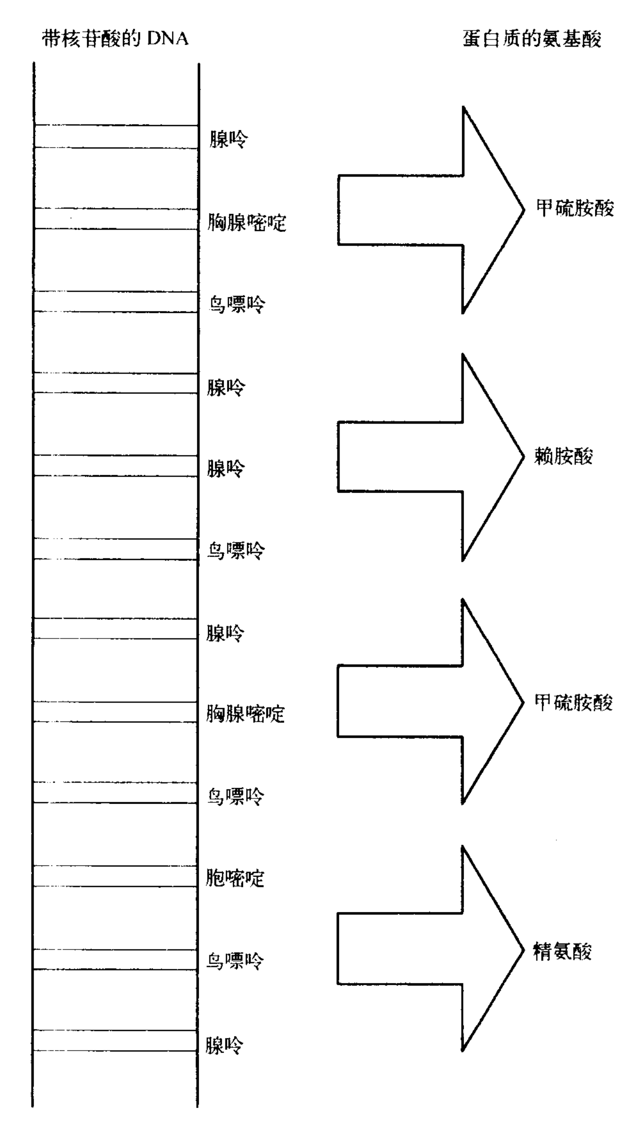
序列含有下列的字母：

ATGAAGATGCGA

那么它解码为一种含有四类氨基酸的蛋白质：甲硫胺酸 ATG、赖胺酸 AAG、甲硫胺酸 ATG 以及精胺酸 CGA（见图）。这种描述故意掩盖了将蓝图变成蛋白质的实际的生物化学机制的细节，但是提供了从 DNA 遗传信息到蛋白质体的一个高层轮廓的映射。

一个生物学编码的例子是从 DNA 的核苷酸到蛋白质中的氨基酸的映射。

在这个简化的模型中，进化过程如下：由 DNA 表示法产生的蛋白质表示为一些生物特征，像蓝眼睛、五个手指、脑的结构、长的象鼻子等等。基因可能表现为一种损坏的方式，导致产生的生物体死亡。健康的生物体生存下去并繁衍子孙，而且把他们的 DNA 传给后代。在高级动物中，DNA 实际上在性别复制期间使用称为交叉的技术，与来自另一位生存者的 DNA 结合。有时，在从一代到下一代的基因传递过程中可能出现一些错误，这就是变异。在许多代之间，所有这些过程的结合使得生物体高度适应环境。这就是进化的过程。



1. 选择

选择类似自然选择过程，在群体中，只有最适应的个体成功地将遗传物质传递给下一代。尽管不像自然界，群体的规模从一代到下一代仍然保持不变，因此没有群体灭绝的机会（这显然不是最优解！）。基因组保持到下一代的机会与它的适应度成比例——与其他的基因组相比，适应度越高，保留到下一代的复制越多。表 13-2 展示了四个基因组与群体适应度的比率。这个比率决定了在下一代中预期的基因组复制数。

表 13-2 使用适应度作为选择

基因组	群体适应度	期望总适应度百分比	复制数
10110	198	40.4%	1.62
00011	84	17.1%	0.69
00010	58	11.8%	0.47
11001	150	30.6%	1.22

预期的复制数是一个小部分，但是群体的基因组数从来都不是微少的。生存是以随机的方式选择与适应度成比例的基因组为基础。我们使用介于 0 到 1 之间的一个随机数，决定基因组的一份复制是否生存。使用表 13-2 的例子，如果第一个随机数小于 0.404，那么选择基因组 10110；如果是在 0.404 和 0.576（40.4% + 17.1%）之间，选择基因组 00011，依此类推。在基因组达到一个合适的数目之前，会产生更多的随机数。使用随机数产生器将该部分概率转换成近似的整数，而且也允许一些低适应度的基因组生存。

在初始的四个基因组上应用选择，产生表 13-3 所示的生存者。注意，大体上，这个过程产生更适应的基因组的多份复制，而产生不太适应者的少量复制。不太适应者 00011，没有平安渡过这一回合的选择，但是最适应者 10110 有两份复制，并且群体的平均适应度已经从 122.5 增加到 151.0。

表 13-3 选择后的群体

基因组	P	适应度
10110	22	198
11001	25	150
00010	2	58
10110	22	198

2. 交叉

应用于生存基因组的下一个操作是交叉。在自然界中发生的交叉，通过把现有的两个基因组的每一块粘贴到一起产生两个新的基因组。如图 13-2 所示，交叉在两个基因组中一个任意的位置开始，第一个基因组的第一部分与第二个基因组的第一部分交叉互换。举例来说，比如两个基因组 10110 和 00010 使用第二和第三位之间的位置进行交叉的情况如下：

10|110

00|010

交叉结果是（来自第二个基因组的基因被划线）：

10|<u>010</u>

<u>00</u>|110

产生的基因组称为孩子，每个孩子都有从双亲继承的一部分染色体。通过选择基因组对，并且掷一枚硬币决定它们是分离还是交叉，然后将交叉应用到群体。这个概率是交叉概率，用 *Pc* 表示。如果它们确实交叉，那么选择一个随机的位置，而且初始基因组的孩子在下代中代替它们。交叉概率的值是 0.5（与投掷硬币相对应）时通常产生好的结果。在该例中，选择两个基因组 10110 和 00010 进行交叉，而且交叉位置是在第二和第三基因（表 13-4）之间。注意在选择和交叉之后，群体的适应度已经从 122.5 升到 183.0，这是经过一代之后一个显著的改进。

表 13-4　选择和交叉后的群体

基　因　组	P	适　应　度
10010	18	234
11001	25	150
00110	6	150
10110	22	198

3. 变异

最后的操作是变异。变异很少在自然界发生，变异是从双亲传给孩子的基因物质被错误编码的结果。因而发生在基因方面的改变可能代表现有群体的适应度的显著改变，虽然结果时常是有害的。选择和交叉在寻找可能的基因组方面效果很好，但是依赖于初始条件和随机性，两者结合可以防止在下一代不考虑特定的有价值的结合。变异提供附加输入。变异率在自然界中相当小，而且对遗传算法来说通常保持相当低——每代的变异大约不超过一个合理的界线。对于刚才的例子，当变异发生的时候，位元从 0 变到 1，或者从 1 变到 0。

假设在这一代中有一个变异，发生在第二基因组的位置 3。表 13-5 展示了这样的一次变异之后的基因组群体。注意，像许多变异一样，这一变异是破坏性的：受变异影响的基因组的适应度从 150 减少到 58，群体的平均适应度从 183.0 减少到 160.0，而且产生的基因组不可能存活到下一代。这是正常的。遗传算法最初的操作方法是选择和交叉。变异具有特别的次级效应，有助于避免未成熟的、局部最适应状态的收敛。当初始群体提供好的可能的组合空间的覆盖时，通过选择和交叉，下一代向最优解快速移动。变异引入的变化可能是毁灭性的，其持续效力不超过一代或者两代。然而，尽管在本例中是有害变异，第二代在初始群体基础上还是有显著的改善。

表 13-5　选择、交叉和变异后的群体

基　因　组	P	适　应　度
10010	18	234
11101	29	58
00110	6	150
10110	22	198

遗传算法的基本原理是，当基因从一代传递到下一代时，通过选择、交叉和变异，持续不断地改进群体的适应度。在特定多代之后——典型的是数十或百代——群体进化接近最优解。遗传算法不总是产生精确的最优解，但是能够非常好地接近最优解。在数据挖掘中，精确的方案未必可行，接近最优解仍然可以产生可操作的结果。

13.1.2 表示数据

前面的例子阐明了将遗传算法应用到简单函数 $31p-p^2$ 的优化的基本机制。该例尝试取一个特殊函数的最大值，函数本身作为适应度函数。基因组相当容易产生，因为函数有一个参数，是5个位元表示的取值介于0和31之间的整数。基因组包含一个单一的基因代表该参数，并且由5个二进制位的序列构成。选择二进制序列表示法不是偶然的。正如本节稍后关于模式的介绍所述，遗传算法在数据的二进制表示（一种非常方便的环境）上效果最佳，因为计算机本身在二进制数据上工作效率最高。

遗传算法不同于其他的数据挖掘和优化技术，它们操纵基因组的位模式，而且一点也不关心有关用二进制位表示的数值，只有适应度函数知道模式的真正意义。适应度函数需要一种能力，把任何基因组转变成一个适应度数值。因为计算机习惯于以位元方式处理数据，所以这一需求似乎不是特别费力。然而，一些位模式可能违犯施加于这个问题上的约束。当基因组违犯这些约束时，适应度就被设为一个最小值。也即，适应度函数的约束测试把约束编入解决方案。

举例来说，前面的例子有一个约束，即数值介于0和31之间。通过使用5个位元表示基因组，就隐含着本约束为真。如果有8个位元呢？在这种情况下，适应度函数看起来像：

- $31p-p^2$，当 $0\leqslant p\leqslant 31$
- 否则为0

这里一般的规则是，对任何没有意义或者违犯问题约束的位模式，设定一个最小适应度数值。这样的模式可能不在初始群体中，但是由于交叉和变异也可能出现。

提示： 适应度函数是定义在以位元序列表示的基因组上，能够理解1和0组成的任何位元序列。当一个特定的位元模式没有一点意义的时候，适应度函数应该返回一个非常低的数值，因此模式不会传给下一代。

13.2 案例研究：使用遗传算法进行资源优化

遗传算法已被证明相当成功的一个领域是，带有一系列约束的资源调度（scheduling resource）问题。这类问题包括有限资源的争用，遵守描述关系的一组复杂的规则。这类问题的关键，是定义一个适应度函数，将所有的约束纳入一个单一适应度数值。这些问题已经超出本书讨论的传统数据挖掘问题的范围；然而，它们是重要的，而且显示了遗传算法的功能。

这个问题的一个实例是在一个门诊部，分配40个内科医生到不同的科室，正如Ed Ewen博士在德拉瓦州的医学中心所遇到的情况一样。门诊部每周工作7天，而且医生在全年中被指派为每周的某一天工作，不考虑其他科室。最佳分配要平衡一些不同的目标：

- 门诊部必须总有医生值班；
- 门诊部应该平衡考虑第一年、第二年和第三年的医生；
- 第三年的医生每天诊视8位病人，第二年的医生看6位病人，第一年的医生只看4位病人。

迄今为止，这个问题并不那么复杂。然而，每个医生在医院的某个部门，像重病特别护理病房、肿瘤部门或社区医院中，4个星期轮一次班。这些轮班有一些其他的约束：

- 资深的医生被指派去重病特别护理病房的时候，不必去门诊部，但是所有其他的医生要去；
- 资历较浅的医生被指派去心脏病护理轮班的时候，不必去门诊部，但是所有其他的医生要去；
- 在同一天被指派给门诊部的、来自重病特别护理病房的医生不超过两位；
- 在同一天被指派给门诊部的、来自其他轮班的医生不超过三位。

可能出现的一个问题是，在一个轮班期间，五位医生在某一天被指派给门诊部。在下一个轮班期间，一个资深医生在内科重病特别护理轮班，两个资历较浅者在心脏病护理轮班。现在门诊部只剩下两位医生，这对门诊部业务是不够的。

遗传算法方法认识到，对这个问题或许没有一个完美的解决方案，但是医生在一周的某些天的值班安排或许有一个较好的方案。Ewen 博士认识到，可以使用一个适应度函数捕捉预定计划“状态的优良”。实际上，Ewen 博士使用的函数是一个反适应度函数——数值越高，预定计划越差。这个函数对于违犯约束施加处罚：

- 每天，当门诊部少于三位医生时，增加一个量——量越大，不足也越大；
- 每天，当在门诊部没有资深医生时，增加一个小的量；
- 每天，在轮班时少于三位医生，给适应度函数增加一个大的量；
- 依此类推。

用这些函数建立一个电子数据表，Ewen 博士试着最小化这些函数，以便得到最佳安排。初始安排得分范围在 130 到 140 之间。在几个小时的工作之后，能够将得分减少到 72，已经相当好了。

然而，他利用来自 Ward Systems Group（www.wardsystems.com）公司的、能够嵌入 Excel 电子表格的一个遗传算法软件包，随机地从一个包含 100 个个体的群体进行安排，没有一个是非常好的。在 80 代之后，软件包将得分降低到 21，比用手工方法能够达到的效果好得多。

这个例子给出了可以在优化问题方面应用遗传算法的一种很好的感觉。与大多数的数据挖掘问题不同的是，它们更多是面向规则，而不是面向数据。解决这些问题的关键是将约束纳入一个单一、待优化的适应度函数（通过寻找一个最大值或最小值）。产生的适应度函数可能高度非线性，难于使用其他技术进行优化。正如我们将会看到的，同样的技术适用于具有大量数据的情形。

提示：当在问题中的规则比数据多时，遗传算法是一个好的工具（虽然在其他领域也是有用的）。这种类型的规划问题，时常包括有限资源的争用，趋向于描述资源及其使用者的一系列复杂的关系。

13.3 模式：遗传算法为什么起作用

乍一看，本章前面介绍的选择、交叉和变异机制没有什么神圣的。举例来说，为什么交叉只选择一个中间点，而不是两个或更多？为什么低的变异率产生较好的结果？假如多重交叉点会更快产生更好的结果，或者高的变异率会更有效，那么自然界以这种方式运转的事实就不是充分正确的。

对于解决产生可操作结果的问题，遗传算法已经在实践中很好地发挥作用，这一事实可

能是继续使用它们的充分理由。然而，知道这项技术有一个理论基础是令人鼓舞的。Holland 教授在 20 世纪 70 年代早期发展了模式处理理论，解释为什么选择、交叉和变异在实践中工作良好。即使遗传算法被埋藏正在使用的工具中，我们特别建议对使用遗传算法解决问题感兴趣的读者，去理解模式；因为这种理解解释了这项技术的能力和局限性。

模式（Schema），来自意思是“form”或“figure”的希腊词，仅仅是出现在基因组中的模式的一种表示法。Schemata（复数形式，从希腊词根得来）被表示为符号序列。基因组的 1 和 0（被称为*固定位置*）通过增加一个 * 进行扩张，* 与一个 0 或一个 1 相匹配。模式和基因组之间的关系很简单。当在模式中的固定位置与基因组中的对应位置相匹配时，基因组与一个模式匹配。通过例子可以更清楚地说明这一点。下列模式：

10 * *

与下列所有四个基因组相匹配，因为它们都有四个符号，以 1 开始，后面跟一个 0：

1000

1001

1011

1010

模式的阶（order of a schema）是它含有的固定位置的数目。举例来说，1 * 10111 的阶是 6，* * * 1010 * 1 的阶是 5，0 * * * * * * * * * * * * * * 的阶是 1。模式的*定义距离*（defining length）是最外层的固定位置之间的距离。因此 1 * 10111 的定义距离是 6（从左边数，7 − 1），* * * 1010 * * 1 的定义距离是 6（即 10 − 4），0 * * * * * * * * * * * * * * 的定义距离是 0（即 1 − 1）。

现在，让我们考察以术语模式表示的适应度函数。如果基因组 000 从一代传到下一代，那么模式 0 * * 也已经生存，* 0 *、* * 0、* 00、0 * 0、00 * 和 * * * 也一样。那么特定模式的适应度，是在给定的群体中与模式匹配的所有基因组的平均适应度。举例来说，模式 0 * * 的适应度是基因组 000，001，010 和 011 的平均适应度，因为当这些基因组生存的时候，模式生存，至少只考虑选择操作时是这样。考虑前面使用的适应度函数为 $31p - p^2$ 的例子中，两个模式 10 * * * 和 00 * * *，初始群体的一个基因组与 10 * * * 匹配，因此它的适应度是 176。与 00 * * * 匹配的两个基因组的适应度是 87 和 58。第一个模式比第二个更适应。事实上，在下一代中只有一个基因组与 00 * * * 匹配，有两个与 10 * * * 匹配。更适应的模式已经生存和繁殖；不太适应的正在消失。

用几何学表示模式有时有助于更好地理解这一概念。考虑长度为 3 的八个可能的基因组：000，001，010，011，100，101，110 和 111，将其分布在单位立方体的两个角上，如图 13-3 所示。模式对应于立方体的边和面，边是阶为 2 的模式，面是阶为 1 的模式。遗传算法处理不同的基因组，也处理由立方体的这些特征显现的模式。包括立方体各块的群体尝试找出具有最佳适应度的角，而模式提供关于可能的解决方案的更大区域的信息。这种几何学的观点可以推广到高维，选择、交叉和变异操作与高维空间的一些超立方体切块对应，比较难以显现。

考虑模式 1 * * * 1。在初始群体中这也是相当适应的，其适应度为 150。初始群体中有一个基因组与之匹配，在下一代中也有一个相同的。这个模式生存下来，仅仅因为包含它的基因组与另外的基因组没有交叉。交叉或许会破坏它。下面与交叉之后的 10 * * * 进行比

较。模式的定义距离愈短，或许愈可能从一代生存到下一代。因此，即使非常适应的较长的模式也极有可能被比较短但适应的同辈代替。使用比较复杂的交叉技术，诸如作两个切块，会完全改变其行为。用更复杂的技术，定义距离不再有效，而且 Holland 在模式上的结果不再成立。

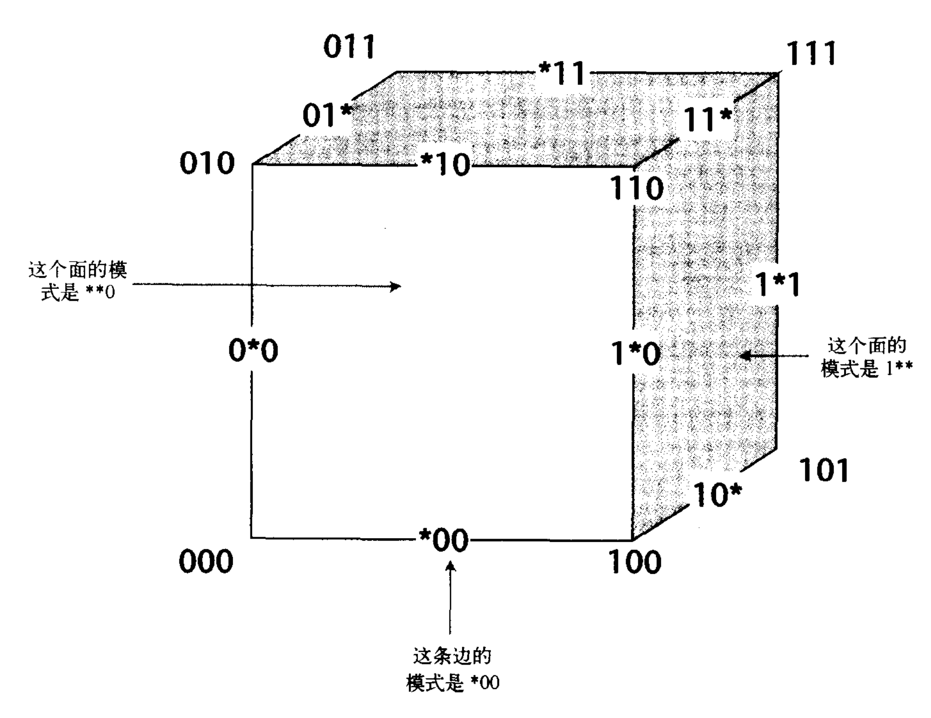


图 13-3 立方体是模式的一种有效的 3 位元表示。角表示基因组，边表示阶为 2 的模式，面表示阶为 1 的模式，整个立方体表示阶为 0 的模式

Holland 精确地证明了这两个观察，并总结为模式定理（也称为遗传算法的基本定理）：短的、低阶的、具有平均适应度的模式在从一代到下一代的群体中有所增加。换言之，短的、低阶的模式是遗传算法工作的构建块。从一代到下一代，最适应的构建块生存下来，彼此结合产生越来越适应的基因组。

模式定理说明遗传算法确实搜遍可能的模式，寻找从一代生存到下一代的适应的构建块。一个自然的问题是，典型地需要处理多少构建块？此处细节我们一带而过，但是 Holland 证明，包含 n 个基因组的群体，其处理模式的数目与 n^3 成比例。这意谓着即使当前只影响 n 个基因组，每代仍需要评估 n^3 个不同的模式。Holland 称这种性质为*隐含并行性*。遗传算法的计算工作与群体的规模成比例，而且在这个工作中，算法有效地处理与 n^3 成比例的若干模式。隐含并行性不能与在工作站的分布式网络上运行的遗传算法可用的显式并行性混淆，或者与在拥有多处理器的计算机上运行该算法时可用的显式并行性混淆。

模式定理揭示了为什么当基因组的表示法中只有两个符号（0 和 1）时效果良好。发现最佳构建块需要处理从一代到下一代的尽可能多的模式。对于两个符号，给定距离为 length 的不同基因组的数目是 2^{length}，不同模式的数目是 3^{length}。概略地，依据单个基因组要处理的独特模式的数目是大约 1.5^{length}。现在，如果在字母表中有更多的符号，如增加 2 和 3，情

况如何呢？现在给定距离的基因组数目是 4^{length}，而且不同模式的数目是 5^{length}（因为星号增加了另外一个符号）。尽管有更多的模式，与给定基因组对应的模式的数目只有 1.25^{length}。当符号的数目增大时，模式的相对数目减少。考察这一点的另一种方法是考虑模式 *00。如果字母表中只有两个字母，那么只有两个基因组 000 和 100 处理这个模式。如果有四个字母，那么有四个基因组：000、100、200 和 300。因为遗传算法尝试使用给定的群体规模找出最佳模式，增加额外的基因组对搜寻没有帮助。

模式是解决方案的构建块，仅使用两个符号允许模式的最大值由给定的群体规模来表示。这些估计并不精确，但是发人深思。更多精确的考察确认了这一结果，即从模式处理的观点来看，两个符号的字母表是最佳的。

13.4 遗传算法的更多应用

遗传算法已经用来解决一些实际问题。本节讨论遗传算法的两个应用，第一是在神经网络方面的应用，其次是在预言性建模方面的应用。

13.4.1 在神经网络方面的应用

神经网络和遗传算法是自然的盟友。遗传算法的强项之一是处理黑箱的能力，即适应度函数可用，但计算细节未知的情况。使用遗传算法训练神经网络是一个好的例子，虽然这种训练方法并不常见。

图 13-4 举例说明一个简单的神经网络，带有三个输入结点、有两个结点的隐藏层和一个单一的输出结点。使网络运转良好的关键是调整边上的权重，以便使输出产生对适当输入的正确答案。第 7 章讨论了结点内函数的性质，以及标准训练算法如何进行。而对目前的讨论，需要做的是对任何给定的一组权重和输入，网络能产生一个输出。权重是实数，有一个包含一组输入和一个相应的正确输出的训练集。

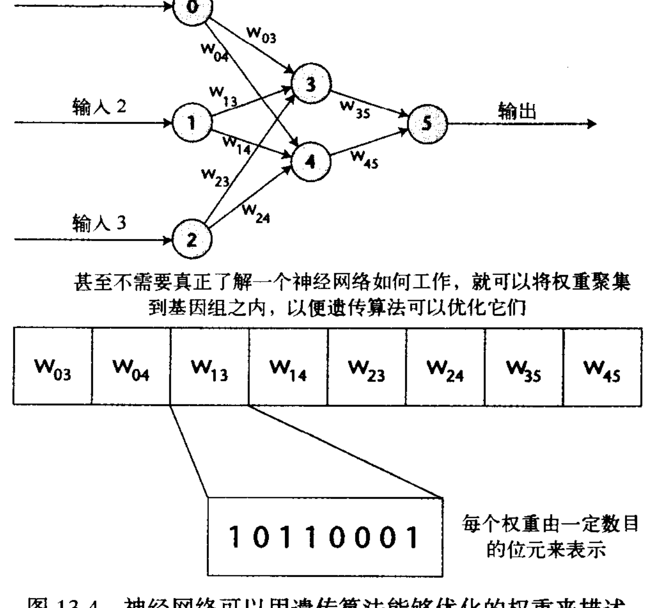


图 13-4 神经网络可以用遗传算法能够优化的权重来描述

第一个需要面对的问题是决定基因组看起来像什么。基因组由所有聚集在网络中的权重构成。适应度函数是什么？适应度函数使用权重产生一个网络，然后把这个模型应用于训练集。然后适应度函数将神经网络的预测输出和实际的输出进行比较；因此，适应度函数被定义为具有训练集上的那些权重的神经网络的全部误差。遗传算法通过最小化这个函数进行。

在神经网络的另一个应用是决定网络的拓扑结构，即在隐藏层（hidden layer）中应该有多少结点和应该使用哪种激活函数。不同的拓扑结构被描述为不同的权重组，然后遗传算法能继续去发现最佳者。在这种情况下，适应度函数产生被基因组描述的网络，然后使用标准方法去训练网络，并把来自最佳网络的误差作为适应度函数。这是遗传算法用于发现复杂问题最优解的一个例子。

13.4.2 案例研究：为响应建模完善一个解决方案

遗传算法的一个更重要的应用是解决真正的商业问题。客户的直接反馈是强有力的商业信息来源。客户提出抱怨的时候，公司有机会通过迅速地修正问题而赢得一个好印象，或者假如太晚，可以设法弥补问题。对于一些公司，像产品制造商，抱怨能提供实际产品使用的日期，即加入制造业和运送日期的一点附加信息。客户抱怨也给公司提供改进工序的机会，以便未来减少不满意的客户。

在我们为移动电话公司创建保持（retention）模型的工作中，已经看到给客服中心打电话的客户比其他客户更忠诚的情形。显然，响应客户表达的需求，尤其当响应迅速和适当的时候，能使客户变得更快乐和更忠诚。在另一家移动电话公司，呼叫客服中心意味着较高的流失率，无疑是由于在呼叫中心长久的等待。

这一案例研究讨论了使用遗传算法的思想，将抱怨分类为抱怨和称赞。

1. 商业环境

一家主要的国际航空公司客户服务部，处理通过以下几种渠道获取的客户评论：

- 在飞机上提供的杂志中含有的响应卡
- 在航空公司网站上的评论表
- 给客户服务中心打电话
- 卡、信件和电子邮件消息

不同的评论有不同的响应优先次序。举例来说，称赞可能导致一种自动回复类型的消息，如“谢谢你成为忠诚的客户”。另一方面，所有的抱怨至少需要答谢，而且许多抱怨是需要探究到底的行为。公司响应得越早，那么保存或许是有价值的、但不满意的客户的机会越大。

航空公司的人员花费相当多的时间分析客户评论，首先将它们分类为抱怨和其他评论，然后把抱怨送入适当的组进行追踪。当客户已经为丢失行李、航班取消、粗暴的服务或污秽的食物难过的时候，怠慢的或不适当的响应只会使事情变得更坏。这个航空公司决定通过自动操作评论的初始分类来减少对抱怨的响应时间。该方法使用马萨诸塞州 Newburyport 的一家软件公司 Genalytics（www.genalytics.com）的软件来完善解决方案。

2. 数据

无论评论来自哪个获取渠道，所有的客户评论最后都被送到评论数据库中。数据库既包含描述评论的字段，也包含实际的文本。一条完整的客户评论记录有下列字段：

- 日期

- 来源（电子邮件、意见卡、电话联系、书信、其他）
- 航班号
- 服务类别
- 始发机场
- 目标机场
- 里程计数
- 收到评论的部门
- 涉及到的航空公司职员的名字，如果提到了的话
- 自由文本评论

有些记录可能会丢失一些字段的数据。来自呼叫中心的评论往往填写正确，因为呼叫中心的服务生训练有素。然而，如果让客户自己去填写客户意见卡或电子邮件，是不可能填写所有字段的。

第一步是预处理文本。公司预处理评论，修改特定的拼写错误并产生关于内容（当前是“food”那个字吗？当前是“meal”那个字吗？等等）的很多衍生变量（derived variable）。对数据库中的每个字，假如在所有消息中出现的次数超过最低限度值，而且不是“of”或“the”之类的非常普遍的字，就创建衍生变量。一些新的变量用于传达有关评论的元数据，像字节数、包含的不同字的数目，这些变量一起构成评论的表头。不使用评论本身，而是使用各种不同的衍生变量。

3. 数据挖掘的任务：完善一个解决方案

数据挖掘的任务是提出一个模型，把描述每个客户评论的很多变量作为输入，并且以某种方法把它们结合在一起产生一个分类。特定的任务是，基于是否是抱怨来分类评论的特征标识。有几种方法处理这一点，如使用决策树或聚类。但在这种情况下，公司使用遗传算法。

用遗传算法解决问题，需要基因组和一个适应度函数。基因组以预处理的评论为基础，每个评论对应一个基因组。首先，要多增加几个字段用于相互作用的变量，诸如是否都提到baggage 和 JFK，或者是否都提到“food”和“chicken”。表头、元数据和相互作用的变量构成评论的特征标识，如图 13-5 所示。

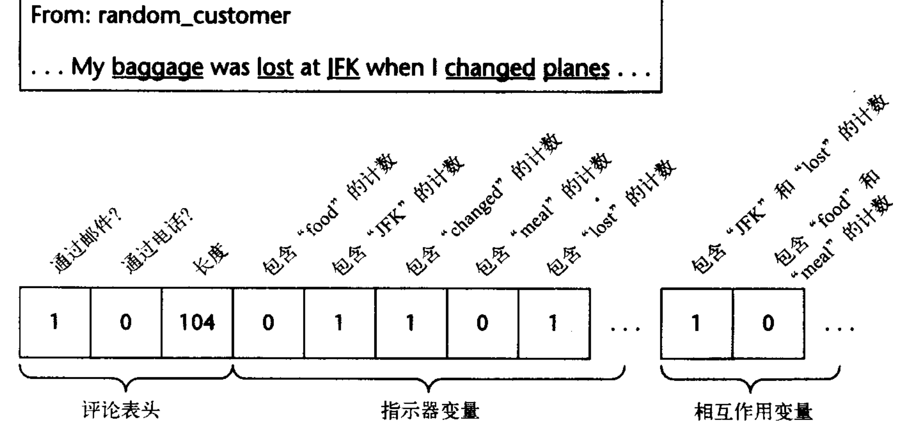


图 13-5 评论的特征标识描述评论文本

评论的特征标识不是基因组，但是与之相关。基因组是与特征标识中的每个变量对应的一组权重（连同一个称为“偏离”的附加权重）。可以将基因组的权重乘以评论的特征标识中的相应字段，以预测评论是否是抱怨，如图 13-6 所示。这是一个单一评论的特征标识的适应度函数。全适应度函数把这一点应用到训练集的所有评论的特征标识。

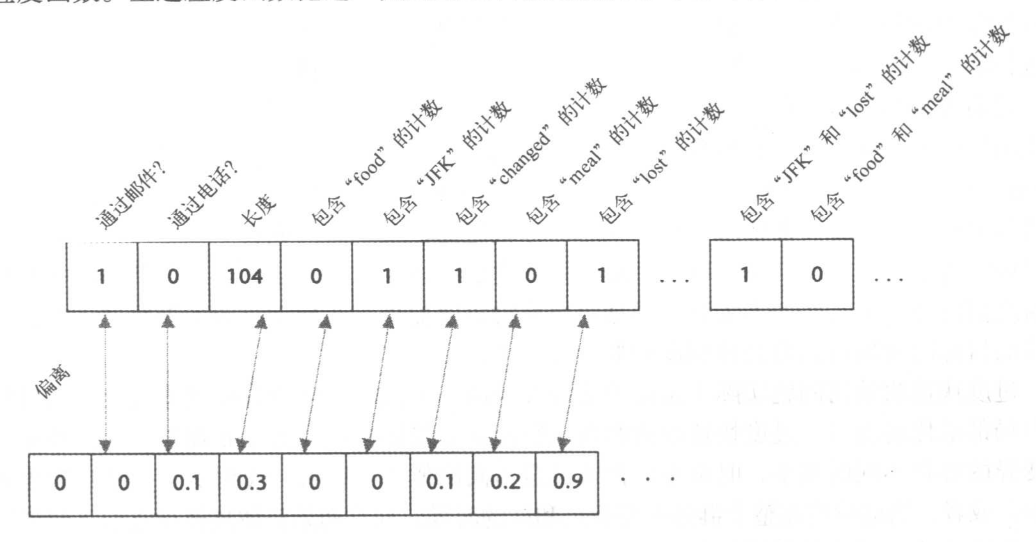


图 13-6 对于评论的特征标识中的每个字段，基因组有一个权重，另外有一个称为“偏离”的附加权重

Genalytics 系统产生一个基因组的随机群体。这些基因组通常把大多数权重设定为低的数值，只把少数一些设定为高的数值。也即，初始群体由评论的特征标识中最简单特征的特定基因组构成。虽然初始群体构成简陋，但是应用选择、交叉和变异后，效果越来越好。在数万代之后，最终模型能够正确分类 85% 的记录，这对加速航空公司的抱怨处理已经足够了。图 13-7 中的图表展示了适应度函数在下一代的改进情况。

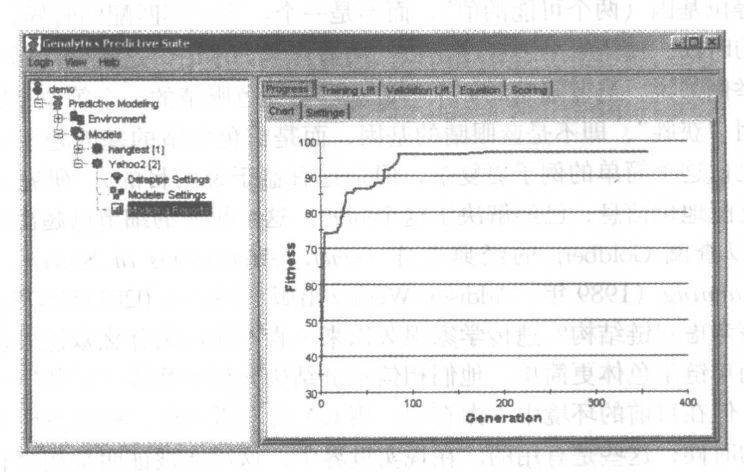


图 13-7 Genalytics 系统展示了训练过程，以及适应度函数如何随每一代有所改进

13.5 超越简单算法

研究人员已经在几个方向上扩大了遗传算法的界限。一些增强是对基本算法的提炼，其他一些修改了基本算法，以便在自然界中提供一个遗传基因活动的更好模型。这项工作时常在机器学习的领域下进行，是当前人工智能的研究领域，目的是使计算机能够模仿人类的方法进行学习。像 Genalytic 的公司已经开始把这些进化技术应用到营销。

先前描述的简单遗传算法在几个方面有改进的空间。算法的无效性之一是，从一代到下一代的进化中，整个群体被更换这一事实。这是对自然界中所发生事情的显著的过度简单化 (oversimplification)。一些研究人员已经开始不再更换整个群体，而是搭建可以按照一定大小增长的重叠的群体，即引入拥挤的概念，以决定哪一个现有的成员应该更换。当纯粹应用的时候，容易造成非常快速的收敛，通常是次优解，因为所有的较不合适的基因组在有机会繁殖之前已经被更换——不太合适的基因组有时也能提供一些帮助。为了处理这一点，用于更换的目标时常取自具有高相似度的群体的子集。

过度快速收敛的问题实际上是简单遗传算法的一个问题，因为寻找整体最优解的目标容易与局部最优解混淆。过度快速收敛时常表明搜索是受限制的。为了处理这一点，对于交叉和变异的各种不同的概率，时常在最初被设定为高的值，然后从一代到下一代的过程中逐渐减少。或者，当适应度在整个群体中变得更和谐的时候，允许初始群体规模在收缩之前增长。

迄今为止，讨论的基因组只由一个基因单链构成。我们在中学不是学了 DNA 是由两个缠绕的链组成的螺旋结构吗？而且那些隐藏在过去的中学背景的其他概念，像隐性基因和支配基因，又如何呢？迄今为止使用的遗传学，是以自然界中发现的最简单的染色体为基础，是单线或**单链染色体**。这些染色体容易在不复杂的、单细胞生物体中发现。在比较复杂的生物体中，染色体是双线，或**双链**，正如人类的 DNA 一样。

双链染色体的算法特征与单链染色体大致相同，因为双链染色体可视为两个染色体绑在一起。实际的算法也以非常相似的方式进行：选择、交叉和变异都相同。差别在于，每个遗传基因有两个等位基因（两个可能的值），而不是一个。当它们匹配的时候，没有任何问题。当它们不匹配的时候，使用哪个适应度函数？用遗传学的术语，这是询问哪一个等位基因**处于支配地位**。举例来说，当蓝眼睛的一个等位基因与棕色眼睛的一个等位基因配对时，棕色眼睛的等位基因"获胜"，即不是蓝眼睛的基因，而是棕色眼睛的基因处于支配地位（实际上，眼睛的颜色比这个简单的例子更复杂，但是这有益于说明目的）。研究人员通过包括关于等位基因的支配地位信息，已经解决了这个问题。这个机制的细节已超出本书的范围，感兴趣的读者可以查阅 Goldberg 的经典著作 *Genetic Algorithms in Search, Optimization, and Machine Learning*（1989 年，Addison-Wesley 出版，ISBN：0201157675）。

为什么应该考虑双链结构？遗传学家很久以来一直不明白为什么双链染色体在自然界中居主导地位，而单链染色体更简单。他们相信双链结构允许生物体"记住"一个基因在另外的环境中有效，但在目前的环境中不太有用。用 GA 的术语表达，就是当环境或者适应度函数随时间变化的时候，这些是有用的。在现实世界中，这可能被证明是相当有用的。改变适应度函数的一个例子是确定证券按时间变化的价格函数。给定的证券价格的优势依赖于算法不能控制的因素，像通货膨胀率。"适应度"函数可以通过结合通货膨胀随时间变化的估计来考虑这一点。

13.6 小结

遗传算法是非常强有力的优化技术。优化不是数据挖掘的核心，但是能解决有趣的、重要的问题。事实上，像神经网络这样的一些数据挖掘算法在神秘面纱的背后依赖于优化。

遗传算法功能的关键是它们只依赖于两件事情。第一是基因组，第二是适应度函数。适应度函数从看似位元的一个随机集合产生一个数值，以此使基因组有意义。基因组将问题编码，时常是由状态相等的一组权重构成。遗传算法在各种适应度函数上工作，使得把许多不容易处理的、不同类型的问题进行编码成为可能。

进化的过程从一个随机的群体开始，然后应用三种变换步骤。第一是选择，这意味着从一代到下一代，更适应的基因组能生存下来。这与自然选择相对应。其次是交叉，其中两个基因组交叉片段，也与自然过程相似。第三是变异，其中一些数值被随意改变。不管在自然界还是遗传算法中，变异通常都相当罕见。

应用这三个过程产生一个新代，其平均适应度应该比初始者更大。创建了越来越多的代之后，群体移向一个最优解。这些过程有一个基于模式的理论基础，它解释了遗传算法如何向一个解移动。

遗传算法已经被应用于一些实际问题，常常应用于资源优化问题。然而，正如在对航空公司的评论分类案例研究中介绍的，它们甚至可以用于预言性建模和分类。

第 14 章 数据挖掘贯穿客户生存周期

数据挖掘的目的是帮助商业理解其最重要资产的价值，这种资产就是客户。前几章已经讨论了使数据挖掘成功的一些算法和方法论。本章从特定的技术转向客户。下面的三章继续讨论这一主题，远离技术上的算法，讨论数据和使用数据挖掘需要的系统环境。

几乎对于任何商业，客户都是最重要的资产。然而，因为随时间变更的不同关系的广泛多样性，他们是难以捉摸的。不同的行业有不同的客户定义。在一种行业中，不同的竞争者用不同的方法管理这些关系。一些行业关注服务质量，一些关注便利性，一些关注价格，还有一些关注关系的其他方面。没有两个商业有完全相同的客户定义，也不会在客户关系中以完全相同的方式对待客户。

数据挖掘的目的是补充其他的客户服务，而不是代替它们。企业通过很多渠道与客户交互，如直接邮寄物品、通过呼叫中心、面对面和通过广告。现在，“鼠标加水泥”（click and mortar）的企业经营方法逐渐成为标准，大多数企业为客户提供在线界面。Web，由于具备与客户交互的新能力，有潜力提供丰富的客户行为数据，这些数据可以变成客户关系的新窗口。在很大程度上，认为代替人与人交互的科技能够使得公司更人性化地对待客户，是带有嘲弄意味的。

这把我们带回客户和客户生存周期。本章努力使数据挖掘关注在中心的客户。它首先概要讨论不同类型的客户关系，然后讨论客户生存周期的细节，因为这与数据挖掘有关。本章提供了不同行业中客户关系的定义，以及在决定客户关系何时开始、何时结束时的一些议题。焦点是客户以及客户与公司之间的事务关系。

14.1 客户关系层次

数据挖掘的主要目标之一，是了解客户以及客户与企业之间的关系。更好地了解他们的一个好的开端是使用不同的客户关系层次，了解客户透过行为所暗示的事情。

客户可以产生丰富的行为信息。每一笔支付、到客户服务中心的每个呼叫、在 Web 上的每次点击、每笔交易，都提供有关客户行为的信息，以及何时和哪种干预起作用，哪种不起作用。Web 是一个特别丰富的信息来源。CNN 不知道谁在关注他们的电视新闻节目，《纽约时报》(New York Times) 不知道每位读者阅读报纸的哪个部分，尽管在 Web 上，cnn.com 和 nytimes.com 都提供了一个有关读者兴趣的更好指示。按时间把这些信息源连接回相应的个体具有挑战性（更不要说按时间把读者的兴趣与相关广告连接在一起的挑战了）。

不可能同等对待所有的客户，因为一些客户显然比另一些客户更有价值。图 14-1 展示了客户关系的分类差异排序，是从每个关系的投资价值数量考虑的。一些客户值得投入很大精力维系非常深入的密切关系，是人们围绕的中心。其他的客户太多和个体化，不值得维系个体关系。对于这个组，我们需要使用技术使关系变得更亲密。第三组也许是最具挑战性的，因为他们介于有真正的亲密关系和假亲密关系的人之间。这一组时常含有小的商业关系和间接关系。后面“没有客户关系”部分谈论另一种情形，即不理解和不需要理解最终用户的公司。

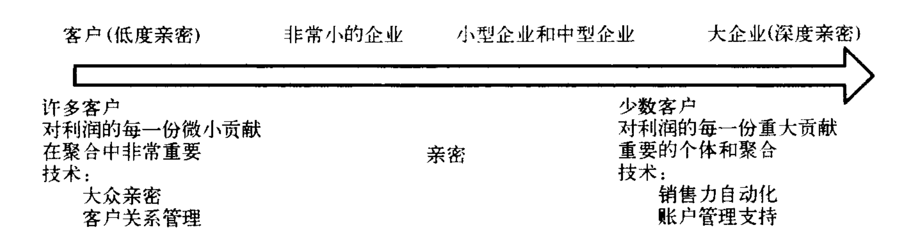


图 14-1　客户关系中的亲密通常随着账户规模的增加而增大

14.1.1　深度亲密

值得维系深度亲密（deep intimacy）关系的客户通常是大企业，即商业客户。这些客户以账户经理和账户梯队的形式出现，大到足以投入专用的资源。这种关系通常是某种企业对企业的关系。一次性的产品和服务刻画了这些关系，使比较不同的客户变得困难，因为每位客户有一组独特的产品。

一个例子是麦当劳公司、可口可乐和迪斯尼三大品牌的强强联手。麦当劳公司是全球最大的可口可乐零售商。当迪斯尼在儿童影院的快餐店进行特别的促销活动时，麦当劳公司首先在欢乐套餐里分发玩具，获得第一份收益。当迪斯尼人物（至少是那些知名人物!）选择汽水或打开冰箱，极有可能就有可口可乐。可口可乐也和迪斯尼有业务合作，因此，迪斯尼在主题公园、旅馆和游乐船上供应可口可乐产品。数以百计的人一起工作，使这三大品牌的联手合作得以顺利进行。数据挖掘，即使拥有在最快的计算机上的最高级算法，也不能代替这些人，这些程序也不会在可预知的未来被自动化。

另一方面，甚至大的账户梯队和个别账户经理都能从分析中受益，特别是使用销售力自动工具。数据挖掘分析通过提供对正在进行的事情的理解，能帮助这样的组更好地工作。数据还能帮助发现一些有用的答案：哪家麦当劳分店特别擅长销售软饮料？产品放置在哪里销售较好？在旅馆和主题公园，天气和饮料消费之间有什么关系？如此等等。

没有客户关系

东京的街道两边有很多排成一行的、像 7-11 的商店，或者在曼哈顿拐角的便利店。这些商店出售小批量的产品，大多是食物，包括刚刚制作的午餐。有三家公司 Lawsons、Seven-Eleven Japan 和 Family Mart 支配这些市场，其中第三大公司 Family Mart 每天的交易额大约 2000 万。日本人口超过 1.2 亿，这意味着平均每个日本人每隔一天要从这些商店之一购买一些东西。这是一种巨大的客户交互量。

下面更深入地考察这些商业活动。这些公司对于它们的客户知道的惟一事情是几乎每一个生活在日本的人至少都是一个偶然的买主，而且在这里几乎全部是现金交易，因此公司无法将客户与在不同商店的一系列时间序列交易联系起来。

这些公司的职能在于分销和付款。在分销方面，他们能够每天给商店送三次货，保证午餐时间的寿司是新鲜的，而且产品没有过期，许多人通过家庭附近的商店用现金支付账单，在现金居统治地位的社会，这是很方便的事情。结合这两种业务情况，一些商店慢慢地变成订单的分段点，客户通过目录或者在 Web 上下订单，然后在舒适的、邻近的便利商店支付和提取货物。

日本的便利商店是一个极端的商业例子，他们对客户了解很少。货物打包厂商是另外一个例子，因为他们不拥有零售关系。

制造商只知道他们何时把货物运给仓库。最终用户的信息仍然很重要，但是其行为数据不在他们的数据库，而在不同零售商的数据库中。为了发现客户行为，他们可能：

- 使用客户的行业范围来发现产品如何被使用
- 通过调查发现客户和他们使用的产品
- 建立和零售商的关系，得到对销售点数据（point-of-sale data）的访问权
- 留意他们正在收集的数据，包括来自 Web、客户服务中心以及邮件的客户反馈信息

分发数据确实有巨大的价值，能够提供正在销售什么物品的线索，以及何时和在哪里销售。其间的潜在信息包括哪则广告消息应该送到哪里，哪一款产品更流行，这都是数据挖掘可以做的事情。

在商务对商务方面，甚至大的财政机构也能从了解客户中受益。世界上最大的一家银行想要分析外汇兑换交易，以便决定哪些客户会受益于以一种货币贷款而使用另一种货币还款，而不是以一种货币贷款并预先兑换还款，目标是为客户提供更好的产品和一个较长期的合作关系。然而，人们需要解释这些结果，依据这些结果行动。

虽然深入的合作关系时常与大型商业主相关，但也不总是这样。零售界的私人银行集团与纯收入高的个体合作，而且给他们提供高度个性化的服务——通常有一个指定的金融家管理他们的关系。当私人银行客户需要贷款或进行投资，只需要呼叫他或她的私人金融家。私人银行集团通常收益颇丰，利润之多得以使他们能不受任何事情的约束。在一家大银行的私人银行集团能够突破公司的信息技术标准，引进 Macintosh 计算机和 AS400，而其他银行的标准是 Windows 和 Unix。私人银行有能力做这件事情，因为他们有经济实力。

同时，仅仅有大型商业主作为客户，并不意味着每个客户都值得如此密切的关注。不管是在 Web 上还是在黄页电话簿上，都有许多商业客户，但是几乎所有的人都被同等对待。虽然客户包含许多大的商业主，但是每个列表带来很少的收入，太少以至于不值得花费更多精力。

14.1.2 大众亲密

另一个极端是大众亲密（mass intimacy）关系。在服务于大众市场的公司，典型地有几十万、数百万或数千万的客户。虽然大多数的客户会喜欢有专门的职员关注自己的需求，但这不是完全经济可行的。公司必须雇用大群人为客户服务，逐渐增加的收益却不能抵消成本消耗。

这是数据挖掘尤其适合客户关系管理（customer relationship management）的地方。许多客户相互作用是完全自动化的，尤其在 Web 上，这具有高度可调整的优点；然而，失去了客户关系管理中的智能化和客户能感受到的温暖感觉。使用技术使关系变得更强大需要多方面的工作：

- 直接为客户工作的人（不管是面对面，通过呼叫中心，或者通过 Web 界面）必须被培训，使其谦恭地对待客户，同时尝试使用增强的客户信息扩展关系。
- 自动化系统（automated system）需要灵活可用，因此可以把不同的消息传递给不同的客户。显然这适应于 Web，但是当获得客户的时候，也适用于账单插页、收银员

收据、后台读取脚本，等等。

- 职员和为客户工作的自动化系统需要能够响应新实践和新消息。有时，这些新的方法来自于职员的良好观念；有时，来自仔细的分析和数据挖掘；有时，来自两者的结合。

这是数据挖掘良性循环（virtuous cycle）的一个扩展。无论是通过算法或人来完成的学习，都需要遵照它行事。产生结果与首先获得它们同样是必需的。成功的事情包括与呼叫中心一起工作，以及培训与客户接触的人。在 Web 上的客户相互作用具有自动化的优点，使得电子地完成良性循环成为可能。人们仍然被包括在管理和确认结果的过程中。然而，Web 使获得数据、分析数据、依据结果行动、不需要离开电子媒介测量结果成为可能。

客户了解的目标可能与有效的渠道操作相冲突。举例来说，美国一家大的移动电话公司，在客户打电话询问与服务相关的问题时，尝试索要客户的 E-mail 地址。有 E-mail 地址有很多好处。一是，未来的服务问题可以通过 Web 处理，花费比通过呼叫中心低。二是，它为偶然的交易消息、交叉销售和保持等机会提供了可能性。然而，因为这个问题在平均呼叫时间中增加了几秒钟，使得呼叫中心业务流量减少。对于呼叫中心来说，得到下一个呼叫比加强与每个客户的关系更重要。

警告： *隐私是主要的关注点，特别是个体客户。然而，对数据挖掘本身是不重要的。在很大程度上，公司之间更关注彼此分享数据，而不是某个公司自己使用数据挖掘了解客户的行为。在法律上，如果把操作目的得来的信息用于像销售或改进客户关系等其他目的，可能是违法的。*

大众亲密提出隐私的议题，这是 Web 发展过程中的主要问题。在研究客户行为的程度上，数据源是在客户和公司之间的转账业务，公司也可以为了像 CRM（虽然对此有一些合法的例外）之类的商业目的使用这些数据，最大的问题在于公司何时出售个体的信息。尽管购买这些数据可能是有用的，或者是有价值的税收来源，但不是数据挖掘必需的部分。

14.1.3 中间关系

中间关系也许最具挑战性。这些客户不够大，不能拥有自己的账户梯队，但也需要特殊的产品和服务。这些可能是中小型的商业主。然而，有一些其他的组，像被称为“大众富裕人”（mass affluent）的银行客户，没有相当的经济能力雇用私人银行，但是仍然需要特别的服务。

这些客户比大众亲密的客户拥有更广泛的产品，或者至少是批量购买享受相关折扣的价格机制，等等。他们也有比较强烈的客户服务需求，有专门的呼叫中心和网站。时常有专门的账务专家同时负责数十个或数百个这种关系。这些专家不总是给予所有的客户同样的关注。数据挖掘应用之一就是传播最佳实践，即发现哪些起作用，哪些不起作用，并且传播这些信息。

在有数万客户的时候，也可以直接使用数据挖掘发现模式，以便从差的客户中区分出好的客户，并且决定下一种产品卖给哪个客户。这种应用和大众亲密是非常相似的。

14.1.4 间接关系

间接关系是另一种类型的客户关系，中间代理促成与最终用户之间的关系。举例来说，保险公司通过代理销售产品，而且通常是由代理建立与客户的关系。有些代理专门销售一家

公司的保单；而有些代理则提供不同公司的分类产品。

这种代理关系带来了商业挑战。举例来说，保险公司一旦与 Data Miners 公司建立一个模型，以确定哪一些投保人可能取消保单。开始这个计划之前，公司认识到实施这样一个计划的严重后果。有了这些信息，代理就可以把高风险的投保人转向其他公司，结果是加速而不是防止这些账户的流失。这家公司并没有实施这项计划。也许问题部分在于了解适当的干预时缺乏想象。公司可以给代理提供特别的激励机制，保持风险客户，对相关的每个人是一种双赢情形。在这种基于代理的关系中，数据挖掘不仅能用于了解客户，也能够用于了解代理。

间接性在其他领域中也有发生。举例来说，信托基金公司通过职员销售退休计划。首要的挑战在于将职员自身包括在该基金中。其次是争取职员报名合适的基金。上述许多健康保险都计划在美国的大公司进行。

产品制造商也有类似的问题。手持电话制造商，像摩托罗拉、诺基亚和爱立信，都想发展一个忠诚的客户基础，因此，经过一代又一代手机，客户继续返回本公司。汽车制造商有相似的目标。制药公司传统地是把药物卖给开处方的医生，而不是直接卖给使用药物的人，虽然像 Viagra 之类的药物在市场上也有销售。产品间接销售活动的另一个好例子是个人计算机中的“内置英特尔”（Intel Inside），即必须为很少有用户见到过的芯片建立商标忠诚度的质量标志。然而，英特尔几乎没有关于拥有这种标识的笔记本电脑的人和公司的信息。

14.2 客户生存周期

客户容易被认为是静态的、不变的实体，他们构成了整个“市场”。然而，这不正确。客户是人（或人的组织），而且他们随时间变化。了解这些变化是数据挖掘价值的一个重要部分。

这些变化称为客户生存周期。事实上，有两个值得注意的客户生存周期，如图 14-2 所示。第一是生存阶段（life stage）。对于一个个体，指的是生存事件，如从中学毕业、有了小孩、找到一份工作，等等。对于一个商业客户，生存周期时常指的是商业的规模和成熟期。第二个客户生存周期是关系本身的生存周期。这两个生存周期彼此相互独立，两者对商业都非常重要。

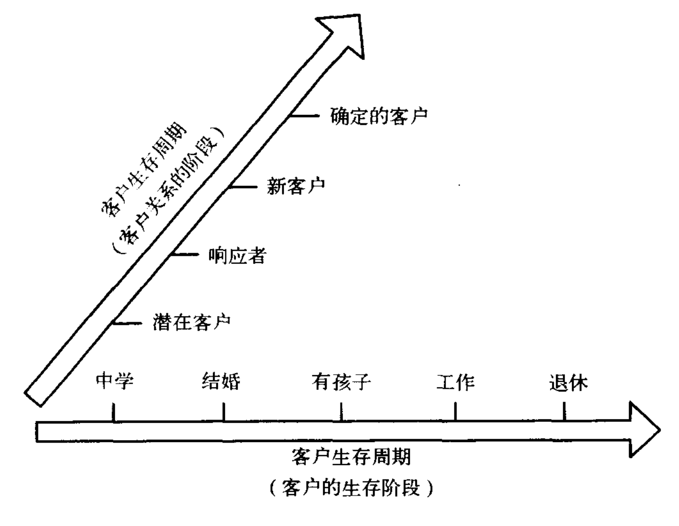


图 14-2 两个客户生存周期

14.2.1　客户生存周期：生存阶段

客户的生存周期由客户关系之外的事件组成，表示个体客户生命的里程碑。这些里程碑由每个人都熟悉的大大小小的事件组成。

客户生存阶段的概念是有用的，因为人，甚至生意人，需要了解这些事件以及它们如何影响个别的客户。举例来说，搬家是一件重要的事件。当人搬家的时候，时常购买新家具，订阅地方报纸，新开一个银行账户，等等。知道哪些人正在搬家有助于把这类个体作为目标，尤其对于家具经销商、报纸和银行。这对许多其他的生存事件也是适用的，从中学毕业到大学、结婚、生孩子、换工作、退休，等等。了解这些生存阶段，公司能够针对特殊群体设计产品和消息。

对于小生意，这不是问题。一家婚礼礼服店专作结婚礼服，对于它们，这样的业务增加，不是因为女人更时常结婚，而是通过推荐。同样地，搬家公司不需要鼓励他们的最近客户重新迁移，需要的是引进新的客户。

另一方面，较大的公司很少专门关注一个生存阶段。他们想要使用生存阶段信息发展产品，而且增强市场营销的针对性，但是有一些新问题。第一，客户的特定环境通常在企业的数据库中不总是可用。一个解决方案是使用购买的信息来扩展数据库。当然，这种扩展数据元素从来不是对每个客户都可用，并且，即使这样的扩展数据在美国可用，也不见得对不同的隐私法律都有效。这种外部数据资源象征过去发生的事情，对当前的生存阶段只是一个推论。

甚至当客户不提供有用的信息时，公司时常会忘记他。举例来说，当客户搬家的时候，会提供新的地址代替旧的地址。但多少公司同时保存这两个地址？并且这类公司中有多少确定客户是正在上移还是下移？通过使用附加人口统计学数据或人口普查数据测量邻近地区的富裕程度？即使有的话，也是很少的。

同样地，许多女人结婚后改变名字，而且把这些信息提供给做生意的公司。在某一点上，两个人结婚后，双方开始合并财产，例如，两人拥有一个活期存款账户，而不是两个。大多数公司不记录客户何时改变了名字，从而失去了提供改变财政环境的目标消息的机会。

在实践中，以生存阶段为基础管理客户关系是困难的：

- 难以用及时的方式识别事件。
- 许多事件是一次性的，或者非常稀有。
- 生存阶段事件通常是不可预知和难以控制的。

无论如何，这些缺点并不使生存阶段没有利用价值，因为生存阶段提供了关于如何用特别的信息联络客户的关键信息。举例来说，广告客户很可能包含不同的信息，这依赖媒体的目标对象。然而，在发展和客户的长期关系的重要性方面，我们想要问：是否有办法改进客户生存周期的使用方法？

14.2.2　客户生存周期

客户生存周期提供另外的了解客户的维度。这尤其关注商业关系，基于客户关系随时间发展这一观察。虽然每宗生意是不同的，但是客户关系把客户归入五个主要的阶段，如图14-3所示：

- 潜在顾客是目标市场中当前还不是客户的人。
- 响应者是已经表现出一些兴趣的潜在顾客，举例来说，填写一个申请表或者在网站上登记。
- 新客户是已经做出承诺的响应者，通常是协议支付，像已经进行第一次购买、签署一份合同或者在网站登记了一些个人信息。
- 确定的客户是那些返回的新客户，对这些客户，希望拓宽或者加深关系。
- 前客户是那些离开的人，包括自发流失（因为他们已经投奔一个竞争者或者不再见到产品的价值）、强制流失（因为没有支付账单）或预期流失（因为不再在目标市场中，如已经搬家）的人。

阶段的精确定义依赖特定的商业环境。举例来说，对于电子媒介网站，潜在客户可能是Web上的任何人；响应者可能是访问过网站的人；新客户是已经注册的人；确定的客户是一个重复的访客；前客户是那些在很长一段时间没有返回的人，这个时间依赖于网站的本质。对其他的商业，这些定义可能截然不同。举例来说，人寿保险公司，有其目标市场，响应者是那些填写一个申请表、然后时常抽血化验的人；新客户是那些被接受的申请者；确定的客户是那些为保险支付保险费的人。

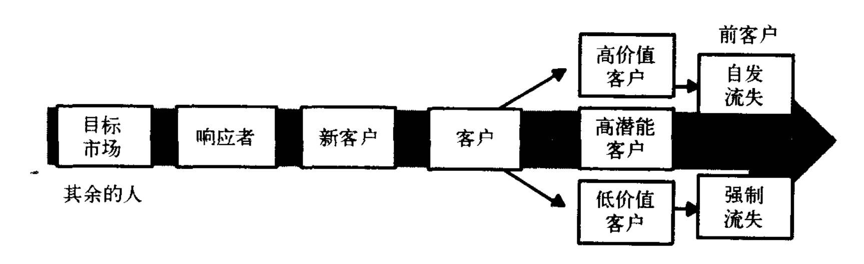


图 14-3 客户生存周期在不同阶段的进展

14.2.3 基于订阅关系和基于事件关系的比较

客户生存关系的另一个维度是交易中固有的承诺。考虑电话用户的下列方式：

- 从投币式公用电话呼叫
- 购买一张数分钟的预付电话卡
- 购买预付费的移动电话
- 选择一个长途电话局
- 购买没有固定条款合同的后付移动电话
- 购买有合同的移动电话

前三个是基于事件关系的例子。后三个是基于订阅关系（subscription-based relationship）的例子。下面两节更详细地探究这些关系的特性。

提示：一个持续的账单关系是持续订阅关系的好信号。这种持续的客户关系提供在商业活动期间参与客户对话的机会。

1. 基于事件的关系

基于事件的关系是在客户部分的一次性承诺。客户可能返回，或者不返回。在上述的例子中，电话公司可能根本没有关于客户的信息，尤其是缴纳现金的客户。这种匿名转账仍然

有信息；然而很明显，几乎没有机会给没有提供联系信息的客户提供定向宣传。

当基于事件的关系成为主流，公司通常通过广泛地传播消息与潜在顾客交流（举例来说，在媒体广告中，免费的固定插页、Web 广告和诸如此类的内容），而不是针对个体发布消息。在这些情况下，分析工作针对产品、地理学和时间，因为这三件事情通常是我们知道的关于客户的交易信息。

当然，广播广告不是惟一接触潜在客户的方法。通过邮件或在 Web 上发放优惠券是另一个方法。在美国的制药公司已经在鼓励潜在客户打电话获得更多的信息，而公司通过这一过程收集呼叫者的一点信息。

有时，基于事件的关系暗示一个与中间人的商务对商务的关系。制药公司在这方面提供一个例子，因为许多市场营销预算都花费在医药供应者身上，公司鼓励他们开某些药物的处方。

2. 基于订阅的关系

基于订阅的关系提供比较自然的了解客户的机会。在前面给出的列表中，后三个例子都有持续的支付关系，其中客户同意随时间的推移支付服务的费用。一个订阅关系提供了未来现金流动（客户未来的付款流量）的机会，以及与每个客户交流的很多机会。

本讨论中，基于订阅的关系是指，按时间与客户保持持续的关系。这可能是支付关系的形式，也可能是零售积分卡或者在网站上注册的形式。

在某些情况下，支付关系是某种订阅，几乎没有提升销售或者交叉销售的余地。因此，已经订购一本杂志的客户可能几乎没有扩大关系的机会。当然，也有一些机会。订阅杂志的客户可能购买礼物订阅或者标有品牌的产品。然而，未来的现金流量在很大程度上由当前的产品成分决定。

在其他情况下，持续关系仅仅是开始。信用卡可能每个月送一个账单；然而，不收费也不欠钱。长途电话公司每个月可能要向客户收取费用，但是可能是月租费。目录公司给客户寄送，但是大部分客户将不进行购买。在这些情况下，消费激励（usage stimulation）是该关系的一个重要部分。

- 基于订阅的关系有两个主要的事件，即关系的开始和结束。当这些事件很明确时，生存分析（见第 12 章）是了解关系持久性的较好候选者。然而，有时定义关系的结束很困难。
- 当客户没有余款，并且在一段特定的时间（如 3 个月或 6 个月）没有转账的时候，信用卡关系可能结束。
- 当客户在一段特定的时间（如 18 个月）没有从目录购买的时候，一个目录关系可能结束。
- 当客户在一段特定的时间（如 12 个月）没使用卡的时候，一个亲和卡关系可能结束。

即使关系相当容易理解，可能也有一些难处理的微妙情形。关系的结束日期就是客户联系或者账户被关闭的日期吗？应该认为没有支付最后账单的客户与因为未付款而被停止的客户相同吗？

这些情形应该作为理解客户关系的指南。值得花时间详细划分客户交互的不同阶段。图 14-4 展示了订阅报纸的客户的不同客户经历。这些客户基本上有以下类型的交互：

- 通过某种渠道开始订阅
- 变更产品（工作日到 7 天，周末到 7 天，7 天到工作日，7 天到周末）

- 延缓递送（典型地是在假期里）
- 抱怨
- 停止订阅（自发或强制的）

在一个基于订阅的关系中，通过收集所有这些不同类型的事件生成客户关系的一张图片，可以在不同时期了解客户。

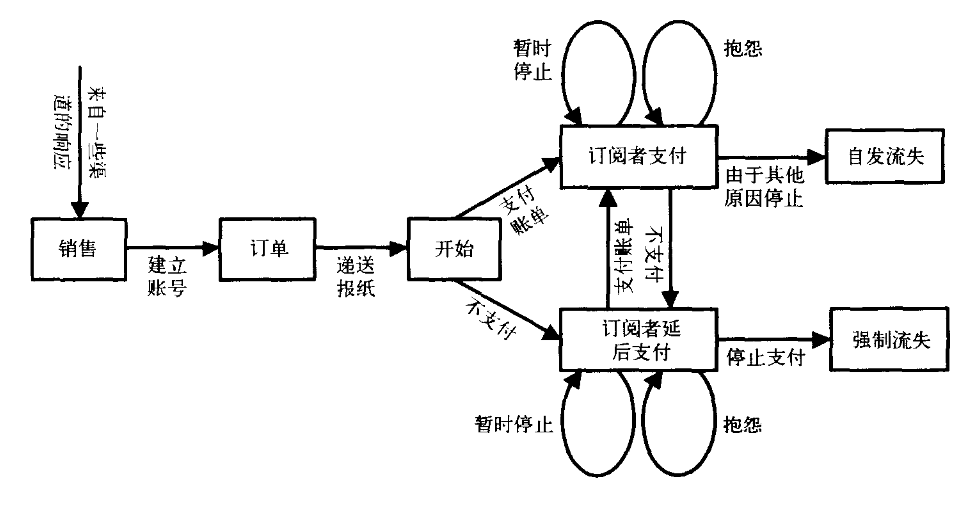


图 14-4　（简化的）订阅报纸的客户的经历，包含一些不同类型的交互

14.3　围绕客户生存周期组织商业过程

客户生存周期以关系的长度和深度等术语描述客户。商业过程使客户从生存周期的一个阶段移到下一个阶段，如图 14-5 所示。审查这些商业过程是有价值的，因为商业的目标之一就是随着时间的过去，使客户变得更有价值。在这一节中，我们考察这些不同的过程和数据挖掘在其中的作用。

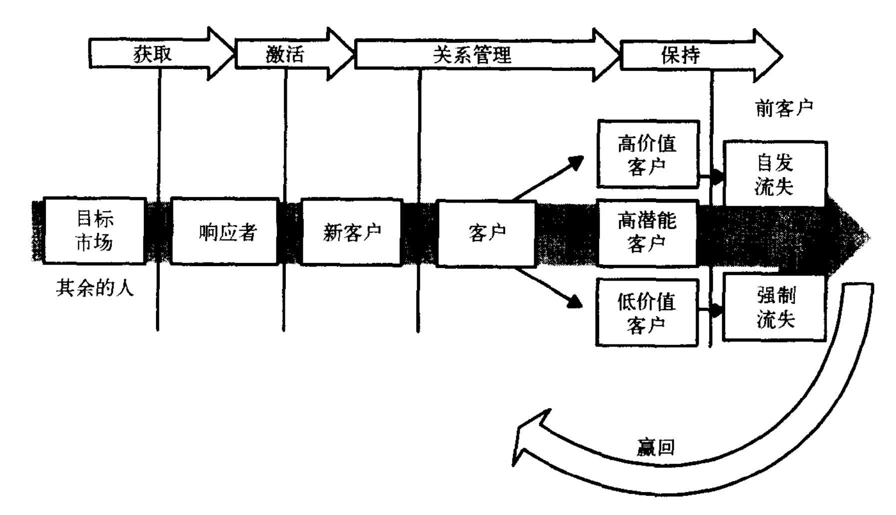


图 14-5　围绕客户生存周期组织商业过程

14.3.1 客户获取

客户获取（acquisition）是吸引潜在客户，并将他们变成客户的过程。通常通过广告和口头消息，以及有目标的市场营销来实现。数据挖掘在获取过程确实能够扮演重要角色。举例来说，第 5 章有一个特别的例子，使用从卡方获得的期望值，突出在不同的区域之中获取的差别。这种描述性分析能够提出在不同地区传播的最佳实践。

关于获取有三个重要的问题，将在本节进行考察：谁是潜在客户？何时获取客户？数据挖掘的角色是什么？

1. 谁是潜在客户

了解谁是潜在客户相当重要，因为宣传应该针对潜在客户。从数据挖掘的观点看，挑战之一是当潜在客户群体改变时使用历史数据。以下是为什么探查潜在客户的时候要特别仔细的三个典型理由：

- 地域扩展带来的潜在客户，可能与原来地区的客户相似或者不相似。
- 产品、服务和定价的改变可能带来不同的目标客户。
- 竞争可能改变潜在客户组合。

一些引发问题的情况是：过去是未来的一个好预言者吗？在大多数情形下，回答应该是“是的”，但是必须巧妙地使用过去。

下列故事是需要当心的一个例子。一家在纽约地区的公司在曼哈顿有一个大客户基础，期待将业务扩大到市郊。他们曾经集中在曼哈顿地区进行直接邮寄营销活动，而且对这次活动的响应者建造了一个模型集。在这个故事中，重要的一个方面是曼哈顿附近富人区的集中度很高，因此，模型集偏向富有者。即，响应者和非响应者比纽约其他地区的一般居民都更富有。

当模型延伸到曼哈顿以外的地区，模型会选取哪些地区呢？它选取的是周围地区中最富有的少数邻近地区，因为这些地区的响应者看起来就像曼哈顿的历史响应者一样。虽然在这些地区有好的潜在顾客，但模型遗漏了许多其他的潜在客户。顺便提一句，这些其他客户通过在邮寄列表中，特别是来自周围地区的名字的随机取样使用对照群组被发现。在对照群组中一些地区有相当高的响应率；尽管是富有的地区，但是不像用来建立模型的曼哈顿邻近地区一样富有。

警告：当把响应模型从一个地理区域扩展到另外的区域时，要特别小心。结果告诉你的可能更多是有关相似的地理特性，而不是响应情况。

2. 何时获取客户

获取客户通常有一个潜在的过程，细节取决于特定的行业，但是一般步骤如下：

- 客户在某一天以某种方式响应。这是“销售”日期。
- 在一个基于账户的关系中，账户被建立。这是“账户开启日期”。
- 账户以某种方式使用。

有时，所有的这些事情同时发生。然而，总是有复杂的因素，如不正确的信用卡号码、错误的拼写地址、买主懊悔，等等。结果可能是有几种日期与获取日期对应。

假设所有的相关日期是有效的，使用哪个最好？那取决于特定的目的。举例来说，在投放一个直接邮件或一个电子邮件之后，就像图 14-6 显示的一样，从响应曲线获知响应者何

时希望加入是有意义的。对于这个目的，销售日期是最重要的日期，因为它表明客户行为，而且问题是有关客户行为。在这个案例中，到底是什么使得开户日期延迟不重要。不同的问题会有不同的答案。举例来说，为比较不同组的响应情况，开户日期可能更重要。登记了"销售"但从未开户的潜在客户应该排除在分析之外。在目标是预测将要开户的客户数的应用中，这也是正确的。

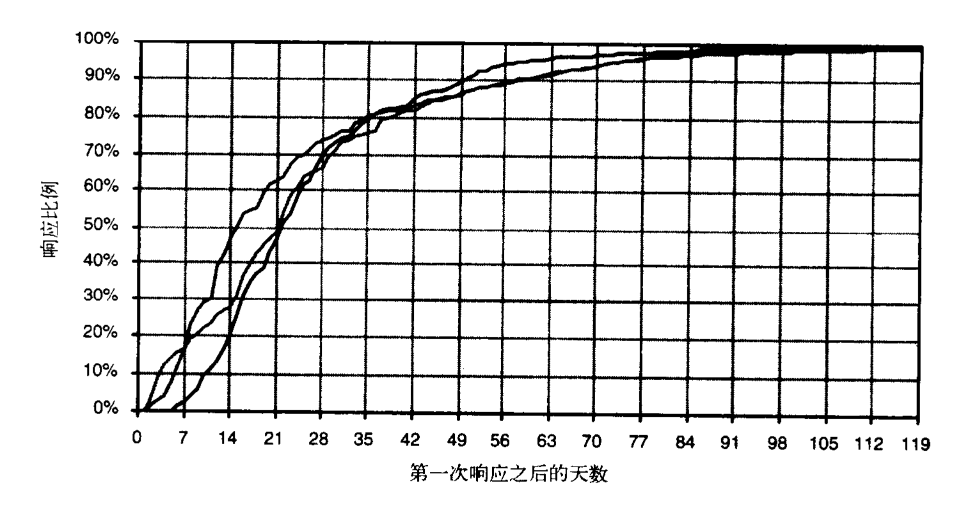


图 14-6 三种直接邮寄营销活动的响应曲线表明，80%的响应来自 5 到 6 个星期内

3. 数据挖掘的角色是什么

可用的数据限制了预言性建模的角色。预言性建模用于像直接邮寄和电话推销类渠道，其中用于联络的花费相对较高。目标是尽可能联系更有可能响应并且成为好客户的潜在客户。对这类工作，可用的数据分为三类：

- 潜在客户来源
- 个体/家庭的附加数据
- 在一个地理层次（典型的户口普查区或户口普查区组）的附加的人口统计学数据

这里的目的是讨论从数据挖掘的观点探察潜在客户。一个好的出发点是使用典型的获取策略大纲。使用直接邮寄或进行电话推销的公司购买客户列表。一些列表在历史上非常好，因此会被完整应用。对于那些来自不太昂贵的列表的名字，当附加的人口统计学在家庭层次是可用的时候，一个模型的集合基于附加的人口统计学；否则，使用在不同的模型集中的邻近地区的人口统计学代替。

定向市场营销的挑战之一是回声效应，即潜在客户通过一种渠道被联络上，但可能通过另一种渠道进入。举例来说，一家公司可能给一群潜在顾客发送一个电子邮件消息。一些响应者不是响应在 Web 上的电子邮件，而是可能打电话给呼叫中心。或者客户可能接收广告信息或直接邮寄，然而通过 Web 网站响应。又或者广告活动可能鼓励同时通过一些不同的渠道响应。图 14-7 展示了回声效应的例子，如传入呼叫和直接邮寄两个渠道的关联所示。另一个挑战是下一节描述的客户激活期间的过滤效应。

警告： *回声效应可能不真实地低估或者高估渠道的效率，因为被一种渠道激发的客户可能被归因于另外的渠道。*

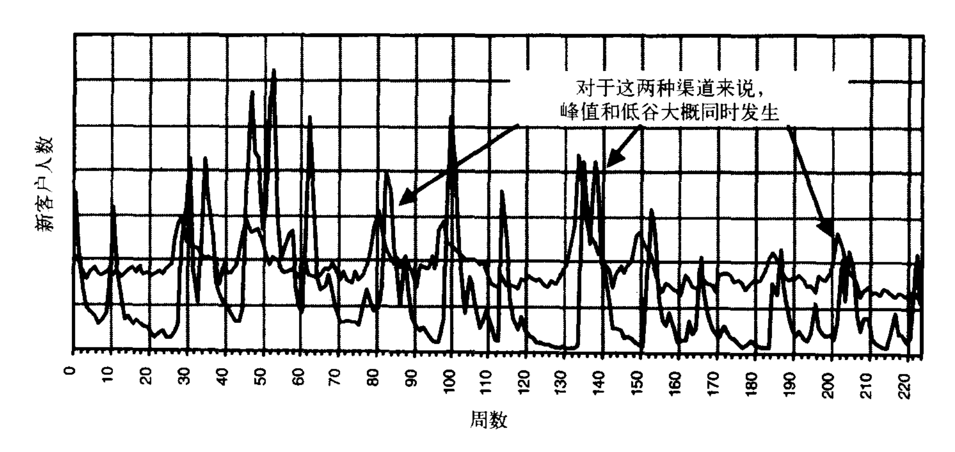


图 14-7 两个渠道之间的关联与时间对比说明，一个渠道可能泄漏进入另一个渠道之内，或者外部因素正影响两个渠道

14.3.2 客户激活

一旦潜在客户表现出兴趣，就有某种激活（activation）过程。这可能像客户在一个网站上填写一个登记表一样简单。或者，可能包括更长的审批过程，像核对信用。或者，可能更麻烦些，如在人寿保险公司的例子中，时常需要进行一项保险业务测试，包括可能在设定等级之前抽取血液样本。大体上，激活是一个操作过程，更多关注的是商业需求，而不是分析需求。

作为一个操作过程，客户激活可能看似与数据挖掘关系不大。但是，有两个非常重要的相互作用。第一是，激活新客户提供了客户在加入时的瞬间状况。这是对客户非常重要的观察，而且作为一个数据来源，也需要保存。初始条件和后来的变化都重要。

提示： *客户激活提供客户关系的初始条件。这种初始条件通常是长期客户行为的有用预报器（predictor）。*

激活也是重要的，因为它缩小并精选客户基。这是一个过滤效应，如图 14-8 所示。这个过程是针对类似报纸订阅等过程的一个常见过程。基本上有下列步骤：

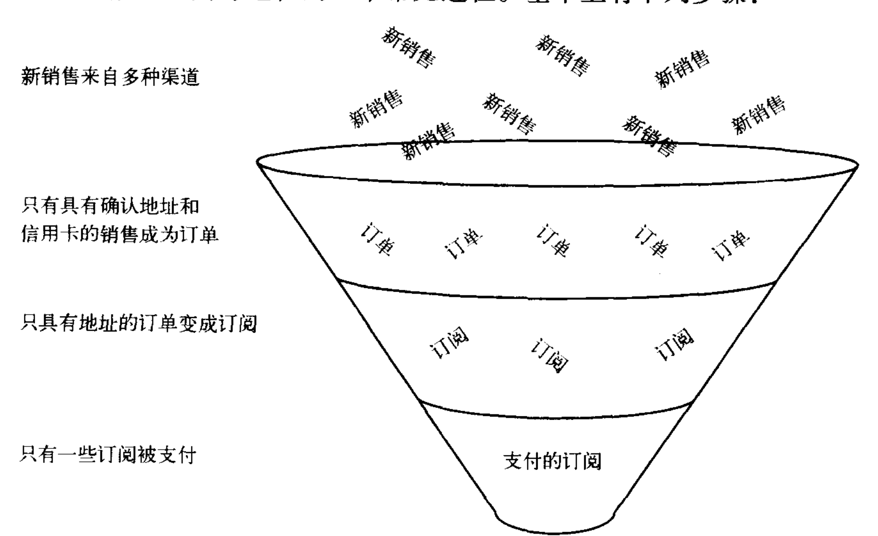


图 14-8 客户激活过程漏斗削减激活过程每一步的响应者

销售：潜在客户表示对订阅感兴趣，通过 Web、电话或投递的响应卡提供地址和付款信息。

订单：建立一个账户，包含关于地址和付款信息的初步确认。

订阅：实际递送报纸，需要进一步确认地址和特别的递送指令。

已支付的订阅：客户为报纸付款。

这些步骤的每一步都失去一些客户，也许只有百分之几或者更多。举例来说，信用卡可能无效，有效期不正确，或者与递送地址不符。客户可能居住在递送区域之外。递送者可能不理解特别的递送指令。地址可能是在一栋不允许进入的公寓大楼内，或者客户根本不支付。这些当中的大部分都是操作因素（客户是否支付是例外），也说明与客户激活有关的操作关注点和过程。

当客户在这个过程中没有以应有的方式移动的时候，数据挖掘能找到原因，并了解在激活期间是什么特征导致客户激活失败。这些结果最大程度地用来改良操作过程，通过强调那些能够带来没有转变为支付订阅的销售的策略，这些结果也提供获取期间的指导。

对基于 Web 的商业，客户激活通常是几乎不需要时间的自动过程，虽然不总是这样。当它很好地起作用的时候，没有任何问题。尽管需要花费一点时间，但却是客户获取过程的必需部分。当它失败的时候，会导致潜在的有价值客户离开。

14.3.3 关系管理

一旦潜在客户变成客户，工作目标是增加客户的价值。通常需要下列活动：

- **提升销售**。让客户购买高级的产品和服务。
- **交叉销售**。扩大客户关系，如让客户购买除了书之外的 CD、机票和汽车等。
- **刺激消费**。确保客户多次回头，举例来说，通过确保客户看到较多的广告，或使用信用卡购买更多的东西。

这三个活动都是数据挖掘所能处理的，尤其是预言性建模，它能够确定对于哪些宣传，哪些客户是最好的目标。这种类型的预言性建模时常确定客户行动的方向，如第 3 章所述。然而，为客户提供适当的营销信息，而不会给他们太多无用的信息，是一项具有挑战性的工作。

电话呼叫和邮件投递是令人烦恼的，不需要的电子邮件消息（时常被称为*垃圾邮件*）在客户关系方面一般具有更消极的效应。一个理由可能是，客户时常为电子邮件支付因特网连接或磁盘空间费用。另一个理由也许是邮件可能在上班时候到达，而不是在家的时候到达，那么就有包含烦人的弹出广告的垃圾邮件问题。当然，这样的电子邮件时常是主动提供的，使那些不想收到诸如赌博诱惑、洗钱、威而刚（Viagra）、性网站、债务还原反应、非法传销模式等信息的人不愉快。

因为电子邮件时常被滥用，即使是公司与客户之间的正常业务沟通也有可能与那些可疑的消息被同等对待。这是一种危险，而且事实上暗示客户联系需要比电子邮件更宽的渠道。

提供许多产品和服务的公司所面临的另一种危险是传播正确的宣传信息。客户未必想要选择；客户只是需要你提供他们想要的。让客户在一大堆市场营销信息中寻找自己感兴趣的东西，不是营销宣传的好方法。因此，有效的方法是对每位客户集中发送少量他们可能感兴趣的产品宣传。当然，每位客户有不同的潜在需求集合。在发现这些关联方面，数据挖掘起

着很关键的作用。

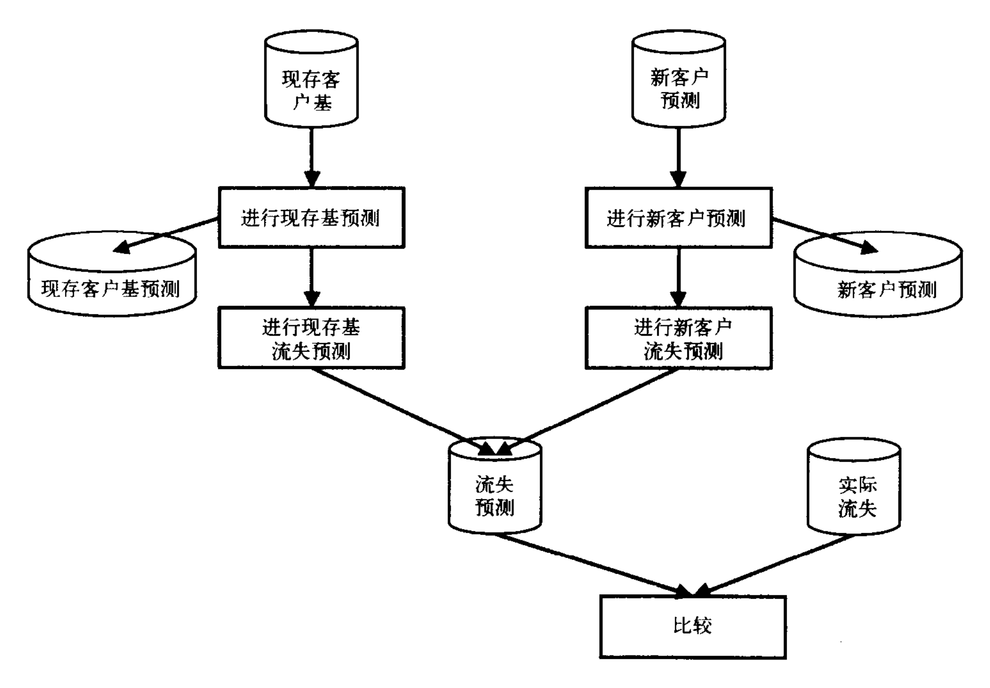
14.3.4 保持

客户保持是预言性建模应用最多的领域之一。考察客户保持有两个方法。第一是在第12章中描述的尝试了解客户保有期的生存分析。生存分析给某一段时间之后可能离开的客户分配一个概率。

流失预测引擎

预测客户停止和客户层次在商业领域起着非常重要的作用，特别是对未来预算规划和市场营销工作。预测提供一个期望值（或一组期望值），用于比较实际发生的事情和期望的事情。这是数据挖掘，特别是生存分析的自然应用。

下列特征展示了预测引擎的外表特征。



预测引擎使用数据挖掘预测客户层次（和流失），同时以背离期望值的方式提供解释

五个重要的输入：

有效日期 该日期之前的所有数字是真实的；而该日期之后的所有数字就是预测。

预测维度 是诸如产品、地域分布和用于发展预测的渠道等客户属性。

新客户 是在有效日期之后被预测维度分解的新客户列表。

活跃的客户 是在有效日期之前活跃的所有客户列表，包括每个客户的预测维度。

实际流失 是分裂成预测维度的实际的停止，用于比较以解释原因。当预测正在进行时，这是不可用的，但是以后就可以用了。

然后预测被分解为下列一些块。现存基预测（The existing base forecast，EBF）决定每个活跃的客户在未来某个给定日期保持活跃的概率，预测直接使用生存分析。新客户预测

(new start forecast, NSF) 决定新客户对未来基的贡献，即这些是在将来仍然活跃的新客户。这是另一种生存分析的直接应用，因为每天都有新客户启动：NSF (t) = One Day Survival of NSF ($t-1$) + New Starts (t)。

流失预测可以容易地从 EBF 和 NSF 获得。现存基流失预测 (EBCF) 是未来在现存基的一个附着块上的流失数目，是生存分析在连续两天的差：EBCF (t) = EBF (t) - EBF ($t+1$)。新客户流失预测 (NSCF) 是未来某一天新客户的流失数目。计算稍微有点技巧，因为必须考虑新客户：NSCF (t) = NSF ($t-1$) - OneDaySurvivalofNSF ($t-1$)。流失预测是这些数目的总和，CF (t) = EBCF (t) + NSCF (t)。

预测的所有块都典型地使用预测维度。结果是预测能与真实值比较，能够用可理解的和对商业有益的术语解释结果。

生存分析的功能在于它关注的时常是保持（客户保有期）的最重要决定因素。通常，已经保有一段时间的客户更可能停留更长的一段时间。然而，通过对几种基本技术的加强，生存分析也能考虑其他一些因素。当有许多数据的时候，使用分层过程可以独立考察不同的因素。当有许多其他因素的时候，参数建模和比例风险（proportional hazard）建模提供类似的能力（本书不详细讨论这些）。在任何情况下，都有可能得到客户的剩余保有期。这不仅对保持干预有用，而且对客户生存值计算和预测客户的数量有用，如上面“流失预测引擎”部分所述。

另一种方法是预测谁在未来的很短时间内会离开。这在很大程度上是一个传统的预言性建模问题，即从过去相似的数据中寻找模式。这个方法对集中的市场营销干预是有用的。知道哪些人在不久的将来会离开，因此使营销活动更集中，投入更多资金挽回每位客户。

14.3.5 赢回

一旦客户已经离开，仍然有可能吸引他们回来。赢回通过提供激励、产品和价格奖励，设法拉回有价值的客户。

赢回倾向于更多地依赖操作策略，而不是数据分析。有时有可能决定客户为什么离开。然而，赢回策略需要作为保持工作本身的一部分开始。举例来说，一些公司特别成立了“挽救梯队”。客户没有和一位专门为保有他们而训练的人交流之前不能离开。除了挽救客户之外，挽救梯队也很好地追踪客户离开的理由——可能是对将来保持客户的工作非常有价值的信息。

数据分析有时能够帮助确定客户为什么离开，特别是当客户抱怨能与操作数据结合的时候。然而，设法吸引不满意的客户回头相当艰难。更重要的工作是设法用具有竞争力的产品、有吸引力的报价和服务，使他们保持第一位置。

14.4 小结

在所有的形式中，客户对商业成功至关重要。一些客户大而且非常重要；这些客户值得特殊对待。其他的客户很小，而且非常多。这是数据挖掘的重点对象，因为始终与每个人保持个人关系代价太高，数据挖掘能帮助提供大众亲密。一些客户介于二者之间，需要在这些方法之间取一个平衡。

基于订阅的关系一般是客户关系的一个好模型，因为这种关系有一个明确的开始和结

束。每位客户有自己的生存周期：婚姻、毕业、生孩子、搬家、换工作，等等。这些对市场营销是有用的，但是问题是，事情发生的时候，公司却不知道。

相反，客户生存周期从商业关系的角度看待客户。首先是潜在客户，被激活会成为新客户。新客户提供了提升销售、交叉销售和刺激消费的机会。最后所有的客户离开，使得保持成为对营销和预测都很重要的一个数据挖掘应用。一旦客户离开，他们可能通过赢回策略被挽回。数据挖掘能提高所有的这些商业机会。

因为世界更多的是被技术驱动，越来越多的数据可用，特别是关于客户行为。数据挖掘意在使用所有这些数据获益，通过汇总数据并在大数据集上应用算法产生意义深长的结果。

然而，在所有这些技术之中，客户关系仍然维持它的中央位置。毕竟，因为是由客户提供收益，客户是商业每年保持成功的惟一秘诀。最后，其他的资金枯竭。没有计算机曾经从 Amazon 进行购买；没有软件曾经在 eBay 上支付一个 Pez 药剂师的费用；没有移动电话曾经预订航班或饭店。总是有人在另一端，不管是个人还是集体。

第 15 章 数据仓库、OLAP 和数据挖掘

自从 20 世纪 60 年代将计算机引入数据处理中心以来，商务活动中的几乎每个操作系统都被计算机化。这些自动化系统管理公司，以自动化方式源源不断地提供大量数据。自动化改变了人们交易的方式和生活方式，ATM（Automated Teller Machine，自动柜员机）、可调节抵押利率、即时库存控制、在线零售、信用卡、Google、24 小时递送、飞行采购社团等就是一些基于计算机自动控制开拓新市场和改革现有市场的实例，这并非是一个新事物，它已经持续了几十年。

在一个典型的公司中，分布有许多各异的系统，从普通的分类账号到自动销售系统，从库存控制到电子数据交换（EDI）等，这些散布的系统创建了大量数据。关于交易活动的特定部分的数据就存在于其中——在某些地方，以某种形式存在。数据是可用的，但没有信息——*即没有在某个确切时间的准确信息*。数据仓库的目的是在确切时间提供确切的可利用的信息。数据仓库就是以决策支持为目的，将贯穿整个组织结构的完全不同的数据集合到一起的处理过程。

数据仓库的作用是作为记录的决策支持（decision-support）系统，使得报告相互一致成为可能，因为它们有相同的潜在的来源。这样的系统不但减少了对全异结果解释的需要，而且在商业企业和时间上提供一致的商务观点。我们相信，随着时间的推移，英明的决策会得到超越时间的更好的底线结果，而且数据仓库帮助管理者做出明智的决策。正如这里所用到的，决策支持是一种有意识的模糊概念，它可以是非常基本的数据，就像每周给予一线管理者的产品报告；也可以是很复杂的，像对潜在客户的深奥建模，使用神经网络去确定提供哪些信息；它也可能（往往是）正好处于上述两种情况之间。

数据仓库与数据挖掘经常是相互关联的，数据挖掘侧重于在数据中发现可操作的(actionable)模式，因此对干净和一致的数据有严格的要求。在数据挖掘背后所做的大量工作往往是识别、获取及清理数据，设计良好的公司数据仓库是一个很有价值的前提条件。更理想的是，如果数据仓库设计中包含对数据挖掘应用的支持，那么这个仓库可以推动和促进数据挖掘工作。同时运用这两种技术是很有价值的。通过把一个干净和一致的不活动数据源转换为可操作信息，数据挖掘完成了数据仓库应该完成的一些工作。

对于这种关系，同样有一些技术成份要求。由于用户同时运行多项工作的能力有限，同时，许多软件，包括数据挖掘和统计学软件，并没有充分利用最快速的服务器上的多处理器和多个硬盘的资源优势，使得数据仓库与数据挖掘不能很好地协同工作。关系数据库管理系统(relational database management systems，RDBMS）是许多数据仓库的核心，是并行处理的，可以充分利用整个系统的资源处理单一查询。更重要的是，用户可能不会注意到这个 RDBMS，因为它的界面没有改变，是 SQL 的一些变形。一个运行在强大的服务器上的数据库，可以成为处理大量数据的强有力的宝贵资源，这种情况就像是在客户层上汇总交易的行为。

数据仓库是非常有用，但这种系统对于数据挖掘和数据分析不是必备条件。统计员、保险精算师和分析师已经使用统计软件包几十年时间——利用它们可以得到很好的结果——他们并没有利用设计良好的中心仓库。然而，由于需要一致、准确和及时的数据来支持商业企业的需要，数据仓库对任何决策支持或信息分析变得日益重要。

本章的重点内容是，把数据仓库作为数据挖掘良性循环（virtuous cycle）的一部分，把

它作为支持该循环的所有四个阶段的有价值的重要组成部分：识别机会、分析数据、应用信息及测量结果。本章的主旨不是指导你如何创建仓库——有很多书籍专注于这个主题，这里真诚地推荐 Ralph Kimba 的 *The Data Warehouse Toolkit*，（Wiley，2002）和 Bill Inmon 的 *Build the Data Warehouse*（Wiley，2002）（中文版《数据仓库》已由机械工业出版社出版）。

本章首先讨论可用的不同数据类型，然后讨论数据挖掘对数据仓库的要求，接着展示了一种典型的数据仓库结构及一些变体。然后转向在线分析处理（OLAP），这是另外一种规范化数据仓库的途径。最后讨论了数据挖掘在这些环境中扮演的角色。不过，像大多数据挖掘相关环境一样，我们首先还是从数据开始。

15.1 数据结构

在计算机上存在许多不同特色的信息，不同层的数据代表不同的抽象类型，如图 15-1 所示。

- 交易数据
- 运行累加数据
- 决策支持累加数据
- 模式
- 元数据
- 商业规则

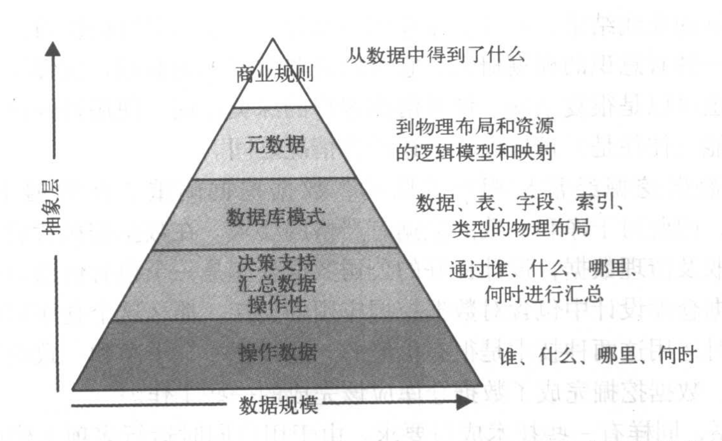


图 15-1 数据分层及其描述帮助用户围绕数据仓库操纵数据。通常数据越抽象，数据量越少

抽象层是数据挖掘所用数据的一个重要特征。在设计良好的系统中，应该可以通过穿越这些不同的抽象层来获取基本数据以支持汇总或商业规则。金字塔的较低层的数据量更大，往往是数据库的素材资料；较高层则数据量较少，往往是计算机程序的内容。所有这些层都很重要，因为我们不希望通过分析详尽的数据而仅仅产生可能已经知道的某些事实。

15.1.1 交易数据——基础层

客户购买的每一种产品、每一次银行交易、每一次 Web 页访问、每一张信用卡的购买、每一个飞行航段、每一个包裹、每一次电话呼叫，在某些操作系统中都被记录下来。每一次，新开设账号或支付账单，在某个地方就应该有一条交易记录，记录了关于谁、什么事情、什么地点、什么时间及花费多少等信息。这种交易层的数据是了解客户行为的原始材料，它是企业的眼睛和耳朵。

不幸的是，随着时间的推移，因为商业变化的需求，操作系统发生了变化，字段可能会随时间改变其含意，这些重要数据只不过被复制和删除。对应于新产品的引入、顾客数目的增加、数据采集、公司重组及新技术采用等各种原因，这种改变时常会发生。考虑数据随时间变化这个事实必须是任何强有力的数据仓库方法的一部分。

提示：数据仓库需要存储数据，以保证信息经历时间的推移仍然是一致的，即使在生产线发生变化、市场发生变化、客户片段发生变化、商业机构发生变化的时候也应该这样。否则，数据挖掘很有可能得到的是反映这些变化的模式，而不是潜在的客户行为。

从交易系统收集的数据量可能是巨大的，一家快餐店在一年时间内会卖出成百上千份快餐；一家连锁超市一天可能会有数万或数十万次交易；一家大银行一天处理数百万张支票和信用卡交易；一个大的 Web 站点每天有数百万次点击（在 2003 年，Google 每天就处理 250 000 000次搜索）；一个电话公司每天有数千万甚至上亿次呼叫；一个大的广告服务器在 Web 上每天追踪超过十亿的广告浏览。即使磁盘价格已经下降，存储所有这些交易信息也还是需要巨额投资的。作为参考，记住一天有 86 400 秒是有意义的，因为一天 100 万次交易实际等于平均每秒 12 次交易（250 000 000 次搜索几乎相当于每秒 3 000 次搜索!）——高峰时期的数值还会更高几倍。

正是因为如此大的数据量，人们通常不愿意在数据仓库中存储交易层数据。从数据挖掘的角度看，这是很遗憾的，因为这些交易是客户行为的最佳描述。

15.1.2 操作汇总数据

操作汇总与交易作用相同，其差别是数据汇总来源于交易。最普通的例子是账单处理系统，它汇总交易数据，通常以每月或每四周为一个周期。这些汇总是面向客户的且常常导致其他交易，如账单支付。在某些情况下，操作汇总可能包括一些字段汇总，这些字段汇总是为了强化公司对客户的了解，而不是为操作目的。比如，第 4 章描述了 AT&T 如何使用呼叫明细记录计算“bizoçity”得分，用于表明一个电话号码出现什么样的呼叫模式时会是类似商务的电话，每次的呼叫记录被丢弃，但得分被更新。

操作汇总数据与交易数据是有区别的，因为汇总是对应于一个时间段，而交易代表每次的事件。让我们考虑订阅客户支付金额数量，在一个账单系统中，支付数量是按照账单定期汇总，有一个付款历史记录表来提供每笔付款交易明细。对于大多数客户来说，按月汇总和付款交易非常相似，但在同一个账单周期内可能出现两笔付款，因而更多的付款明细信息可能对客户付款模式提供有用的深入了解。

15.1.3 决策支持汇总数据

决策支持汇总数据是辅助商业决策的数据。公司的财务数据提供了一个决策支持汇总数据的实例，这通常被认为是决策过程中最清洁的数据。另一个例子是数据仓库和数据集市，其目的是在客户层次上提供记录的决策支持系统。维护决策支持汇总数据就是数据仓库的目的。

通常来说，把同一个系统同时用于分析和操作目的并不是一个好主意，因为操作目的需要优先考虑，这样会产生一个对于操作最优化的系统，这种最优化不是针对决策支持目的的。财务系统一般不是为了解客户过程设计的，因为它们的设计目的是清算账号。把客户汇

总完全结算到分类账号是非常复杂的，而且通常没有必要这样做。数据仓库的目标之一是提供一致的定义和布局，以便相似报告产生相似的结果，无论它们是哪个商务用户生成的或者在什么时间生成的。本章主要关心的是这个抽象层。

从某种意义上说，汇总看似破坏了信息，因为它们对事情进行聚集。由于这个原因，不同的汇总用于不同的目的。销售点交易可以捕获走过扫描器的每一瓶沙丁鱼罐头信息，但只有汇总结果，才能用客户在一天中的习惯购物时间段以及她花在罐装食品部的钱的比例来描述其购买行为。在这种情况下，客户数据汇总看似在创建信息。

警告： 不要期望客户层次的数据仓库信息完全与财务系统数据平衡（尽管这两个系统给出的结果可能很相近）。虽然理论上有可能，但这种平衡被证明是很困难的，往往会使人对数据仓库的目的产生别的分歧。

15.1.4 数据库模式

迄今为止的讨论都是关于数据的。实际上，数据的结构也同样重要——比如什么数据被存储、它被存储在哪里、什么数据不被存储等。后面“什么是关系数据库?”部分解释了这些关系数据库中的主要概念，关系数据库是存储大量数据最常用的系统。

不管数据是如何存储的，区分描述存储的两种方式都是重要的。物理模式从基本软件需要的技术细节上描述它的布局，一个例子就是 SQL 中由“CREATE TABLE”产生的财务报告书；与之相对的是，逻辑模式以一种最终用户更易接受的方式描述该数据。这两种方法不必相同，甚至不必相似，如图 15-2 所示。

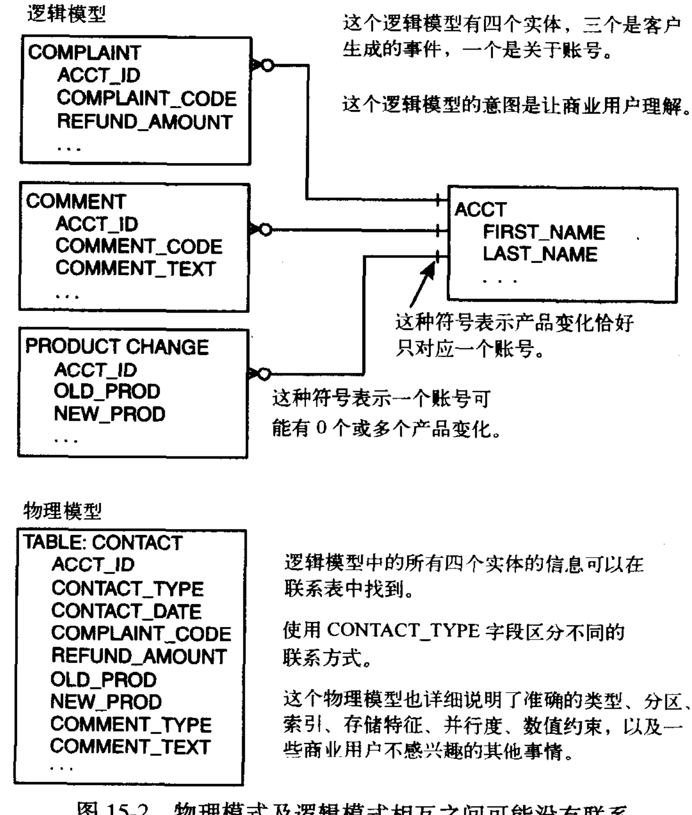


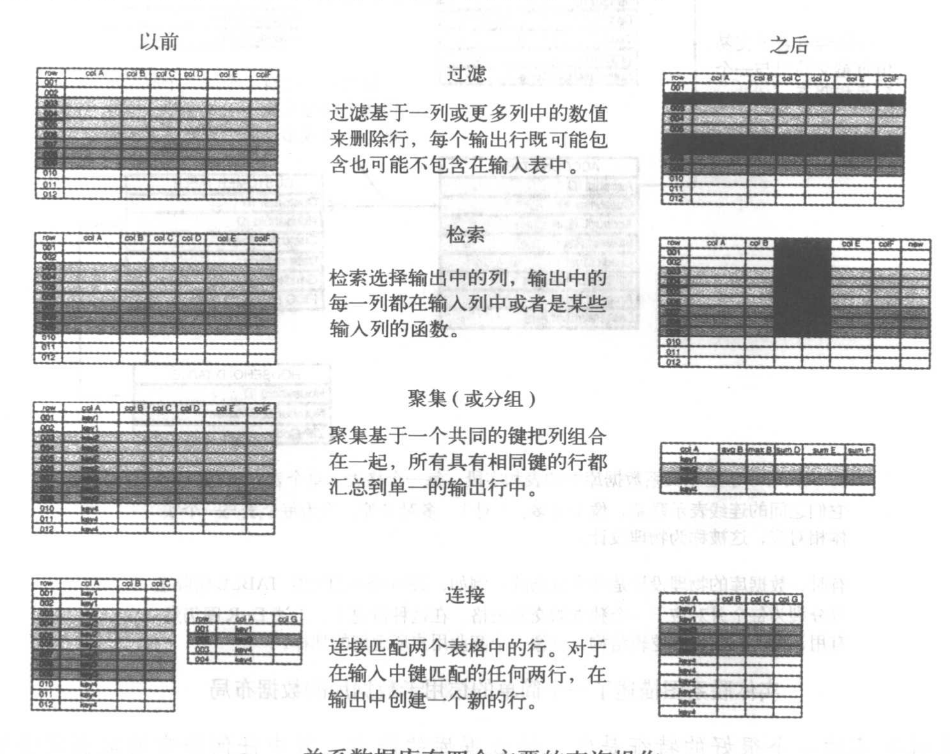
图 15-2 物理模式及逻辑模式相互之间可能没有联系

警告： 数据库中字段的存在并不意味着数据确实是存在的。理解数据挖掘中使用的每一个字段，不能由于字段的存在而假定其定位是正确的，这一点是很重要的，对此存在怀疑总是有必要的。

类比可能会有助于理解物理模式和逻辑模式的作用。逻辑模式描述事物的方式是商务用户熟悉的，比如一个房子是大农场风格，有四个卧室、三个浴室及两个车库；物理模式趋向于对它的结构布局进行更详细的描述，地基是混凝土加固的，有四英尺深，地板面积有1500平方英尺，墙体是混凝土板材等。这种建筑细节资料虽然有用而且完整，却不一定能帮助一个家庭找到合适的房子。

什么是关系数据库？

存储数据最常用的方式之一是把数据存储于关系数据库管理系统（RDBMS）。关系数据库的基本概念始于20世纪70年代早期，是由E.F.Codd在研究由元组（我们在表中称做“行”的东西）构成的一种特殊类型的集合性质开始的。从该研究中，他导出了一个包含运算符的关系代数，形成了关系代数学，下图中描述了其内容。



关系数据库有四个主要的查询操作

这些操作是集合运算（如并集和交集）之外的操作，在非科学技术中，这些关系操作是：

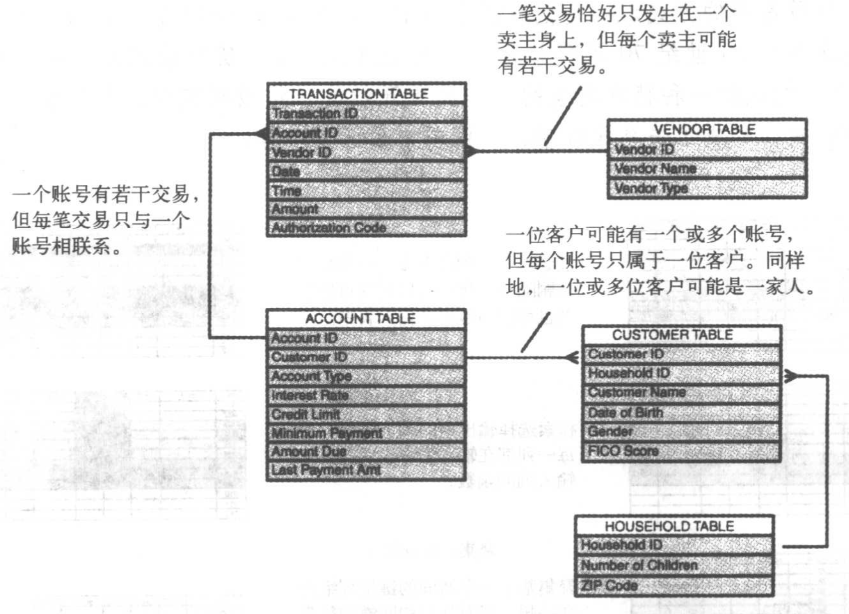
- 基于行中的值过滤一个已知行的集合；
- 检索已知列的集合并对它们进行基本操作；
- 分组行并聚集列的值；
- 基于列中的值将两个表格连接起来。

有趣的是，这种关系操作不包括分类（除非为了输出目的）。这些操作详细说明了对元

组可以做什么，而不是如何去完成。事实上，关系数据库常常利用分类进行分组和连接运算，然而，对于这些运算也存在以无分类为基础的算法。

SQL 是 20 世纪 80 年代由 IBM 开发出来的，它现在已成为访问数据库和实现这些基本运算的标准语言。因为 SQL 支持子查询（也就是说，利用一个查询的结果作为另一个查询的表），这使得表达某些复杂数据操纵成为可能。

表示数据库结构的一种常用方法是使用实体联系（E-R）图，下图是 5 个实体及其四种联系组成的一个简单的 E-R 图。在这个例子中，每个实体对应于一个独立的表，这个表的列对应于实体的属性。此外，还包含一些表示数据库中表之间联系的列，这样的列被称为键（不是外键就是主键）。在数据库表中，使用一致的命名习惯显式地存储键，有利于自如使用数据库。



实体联系图描述了一个简单的信用卡数据库的数据布局

关系数据库的一个很好的特征是它设计数据库的能力，其中任何给定的数据项恰好出现在惟一的位置——没有重复，这样的数据库称为规范化的数据库。理论上讲，准确地知道每一个数据项所处的位置是高效的，因为更新任何字段只需要修改表中的某一行。如果一个规范化的数据库设计良好且得以实现，就没有冗余数据、过时的数据或无效的数据。

规范化背后的一个重要思想就是创建参照表，每个参照表逻辑上对应于一个实体，且都有一个键用以查找关于实体的信息。在一个规范化的数据库中，“连接”操作通常用于在参照表中查找数值。

关系数据库是存储和访问数据的有力方法。然而，这种设计主要关注更新和处理大量的交

易数据。数据挖掘的兴趣在于将数据结合在一起以发现更高层的模式。数据挖掘通常会用到许多查询语句，每一个查询都需要几个连接、几个聚集及子查询——一群真正的杀手查询。

什么是关系数据库?

对于数据挖掘来说，关系数据库（及SQL）有一些局限性。首先，它们对时间序列几乎不提供支持。这使得很难从交易数据中推测出比如第二次产品的购买、客户所响应的最后三次商品促销，或者事件发生的次序等事实，这些可能需要非常复杂的SQL。另一个问题是两个操作常常会无意间消除字段。当一个字段包含一个缺失值（NULL）时，它会自动舍弃任何比较关系，甚至是“不相等”关系。同样地，默认的连接运算（称为内连接）会删除不匹配的行，这意味着客户可能无意间被排除在数据查询之外。SQL中的运算集并非特别充足，尤其是那些文本和日期字段。因此，每个数据库销售商都把标准的SQL进行扩展，以便包括稍有不同的功能集合。

数据库模式同样可以说明数据中不同寻常的发现，例如，我们曾经接手一个美国的呼叫明细记录文件，其中包括以城市和州为字段的每一次呼叫目的地，这个文件包含超过两百个州的代码——比实际的州多出很多。到底发生了什么事情呢？我们发现城市和州字段从来没在操作系统中应用过，因此，它们的内容自然遭到了怀疑——没有用到的数据不大可能是正确的。替代城市和州的是，所有的位置信息都通过邮政编码产生，这些多余的字段是不准确的，原因是，这些州字段被先写上去，而有14个字符长的城市字段被后写上去，于是较长的城市名称覆盖了与之相邻的州字段。所以，“WEST PALM BEACH, FL”最终把“H”放到州字段中，成为“WEST PALM BEAC, HL”的形式，而“COLORADO SPRINGS, CO”则变成了“COLORADO SPRIN, GS”（译者注：这样就会出现很多实际并不存在的州代码，也就是前面提到的美国会出现200个“州”的原因），理解数据分布帮助我们找出这些有趣而不寻常的问题。

15.1.5 元数据

元数据超越数据库模式而给出更多信息，它可以使商业用户理解什么类型的信息被保存在数据库中。本质上讲，它是关于系统的文档编制，包括以下信息：

- 每一个字段许可的数值；
- 对每个字段内容的描述（例如：开始日期到底是销售日期还是激活日期）；
- 数据加载日期；
- 数据最近更新程度的表示（在一个支付周期之后，什么时候支付数据进入系统）；
- 映射到其他系统（某源系统中，表A中的ID就是表B中的ID字段）。

当元数据可用的时候，就提供了一种宝贵的服务；当它们不可用的时候，就需要收集这种类型的信息，通常可以从友好的数据库系统管理员和分析家那里得到——这时每个人的时间效率都较低。对于数据仓库而言，元数据提供规律，因为仓库的变化必然反映在将与用户沟通的元数据中。总的来说，通过让用户更多关注和熟悉数据仓库的内容，一个好的元数据系统有助于确保数据仓库的成功。对数据挖掘者来说，元数据在捕捉和理解数据方面提供了有价值的帮助。

15.1.6 商业规则

抽象的最高层就是商业规则。这些规则描述了为什么会存在关联以及如何应用关联。某些商业规则很容易获取，因为它们表现了商业历史——什么样的交易活动会在什么时候发

生，什么样的产品在何时是有用的，等等；其他类型的规则较难得到，它们经常深埋在代码片段内及旧的备忘录中。没有人会记起为什么欺诈检测（fraud detection）系统会忽略500美元以下的索赔。也许以前有一个好的商业理由，但一旦规则被植入计算机代码，这个理由、这个商业规则便常常会被丢掉。

商业规则与数据挖掘有一个相近的关联。某些数据挖掘技术，像购物篮分析和决策树等，都会产生简明的规则。这些规则通常可能是已知的，例如，获悉电话会议会与呼叫等待一起销售可能没有意义，因为这个特征只是作为捆绑销售的一部分来售出的。或者，一个直接邮寄模型所对应的模型最终目标仅仅是富人区，可能反映出用于建立这个模型的历史数据是有偏离的这样一个事实，因为模型集只在这些地区有响应者。

在数据中发现商业规则既是成功也是失败。找到这些规则是这些复杂深奥算法的成功应用；但是，在数据挖掘中，我们希望找到可操作的模式，然而这样的模式是不可操作的。

15.2 数据仓库的大致结构

有多层途径通往数据仓库，这使我们认识到数据需要有多种不同形式的来源。它提供了一个广泛的系统用于决策支持方面的数据管理，这种结构（见图15-3）的主要组成部分是：

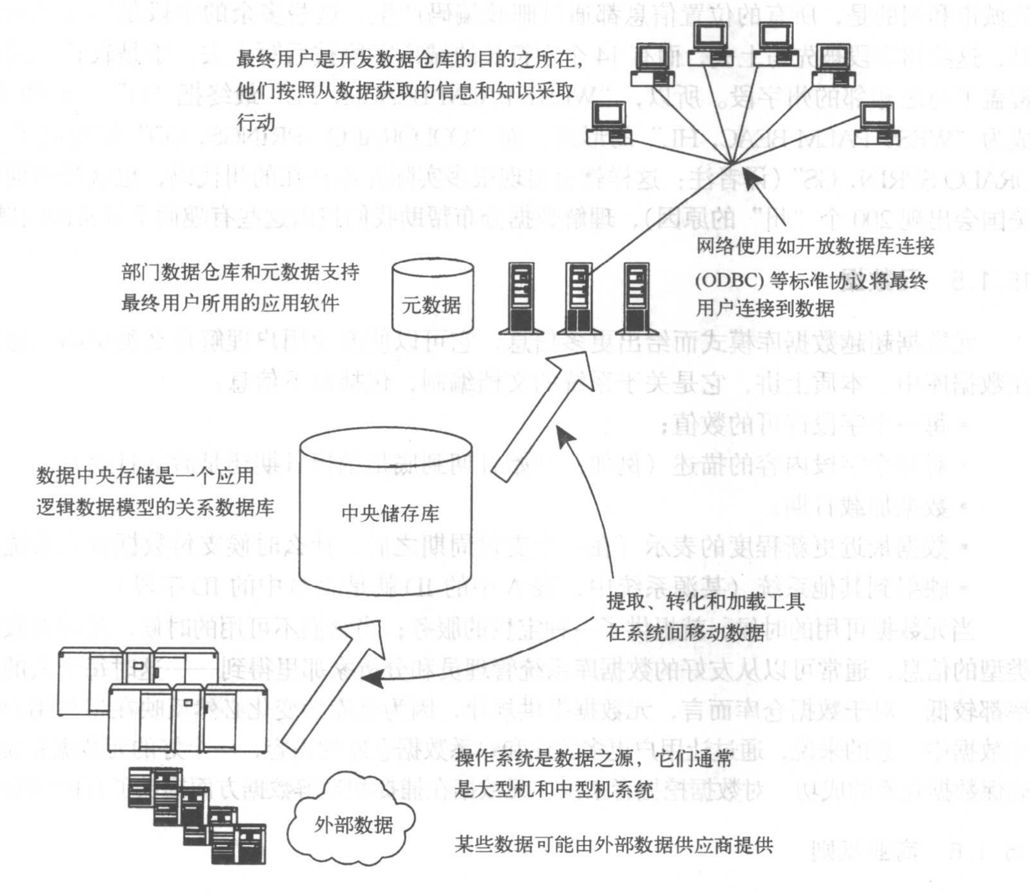


图 15-3 生成数据仓库的多层途径包括中央储存库、数据中心、最终用户工具及将所有这些连接在一起的工具

- 源系统就是数据的来源；
- 提取、转化和加载（ETL）使数据在不同数据存储单元间移动；
- 中央储存库是数据仓库的主存储处；
- 元数据储存库描述哪些数据是有用的以及数据存储在哪里；
- 数据集市给最终用户及应用提供快速、专门的存取；
- 操作反馈整合决策支持，送回操作系统中；
- 最终用户是开发数据仓库的首要原因之所在。

实际上，在每一个系统中都存在一个或多个这样的组成部分，它们是贯穿整个企业决策支持的重要构件，下面关于这些组成部分的讨论将沿着一个数据流途径进行，这些数据像流水一样，它起源于源系统，流过数据仓库的各个组成部分，最后把信息和数值传递给最终用户。这些组成部分依赖于一个由硬件、软件及网络组成的技术基础，这种底层结构功能必须足够强大，以同时满足最终用户的需求，以及不断增长的数据和数据处理的需求。

15.2.1 源系统

数据起源于源系统，源系统通常是操作系统和外部数据输入。这些系统是为了使操作有效而设计的，不是为了决策支持目的，数据反映了这一事实。例如，交易数据可能每几个月清理一次以减少存储压力，同样的信息可能以不同的方式表现出来。例如，一个零售点源系统使用“退还物品”标记来表明退还商品。也就是说，当客户同时购买了一种新商品时除外。在这个实例中，在购买字段应有一个负的数量，在现实世界中这种不规则现象大量存在。

通常，客户关系管理感兴趣的信息不是有意的收集的，例如，以下是从电话公司客户中识别出商业客户的六种可能方式：

- 利用一个客户类型指示项：“B”或“C”对应于商业客户或普通客户；
- 利用费率计划：某些只销售给商业客户，另一些给普通客户；
- 利用获取渠道：某些渠道是为商业客户保留的，另一些给普通客户；
- 利用电话线分机数目：普通客户是1或2，商业客户则更多；
- 利用信用分类：与普通客户相比，商业客户使用的是一个不同的信用卡系列类型；
- 利用一个基于商业客户可能呼叫模式的模型得分。

（显然，这些方式常常并不给出一致的结果。）在数据仓库中面临的一项挑战是获得能够在整个商务活动中使用的一致的定义，做到这一点的关键是元数据能清楚地给出每个字段的准确含义，这样每个使用数据仓库的人都使用同样的语言。

为决策支持收集数据的重点在操作系统，因为这些系统最初是为交易过程设计的。以统一的格式将数据聚集在一起几乎总是实现数据仓库解决方案中花费时间最多的部分。

源系统也带来了其他类型的问题。它们通常运行在各式各样的硬件上，且相当数量的软件是内部创建的或高级用户化。这些系统一般是大型机或中型机系统，且通常使用复杂的、独有的文件结构。大型机系统设计用于支持和处理数据，而不是共享数据。尽管系统越来越开放，访问这些数据将随之成为一个问题，特别是当不同的系统用于支持企业内不同的部门时。而且，系统可能按地理区域分布，这会进一步加剧将数据整合到一起的困难。

15.2.2 提取、转化和加载

通过从源系统将数据映射和移动到其他环境，提取、转化和加载（ETL）工具解决了从

各异的系统收集数据的问题。数据移动和清理以前通常是由程序员负责完成的。必要时，他们编写一段专用的程序代码。当系统扩展以及源系统发生变化时，这种针对特殊应用的代码变得很脆弱。

虽然编写程序可能还是必需的，但现在有些产品已经能解决大部分 ETL 问题，这些工具可以详细列出源系统清单，在不同的表格和文件之间进行映射。它们提供了校验数据的能力，当加载不成功时指出存在的错误，这些工具同样支持在表格中查找数据（所以，只有已知的产品代码可以加载到数据仓库中）。这些工具的目标是描述数据来自哪里，它们出现了什么问题——而不是为了编写出按部就班工作的代码将数据从一个系统提取出来置入另一个系统。标准程序语言，比如 COBOL 和 RPG，关注于每一步而不是需要处理的整体问题；ETL 工具常常提供一个元数据界面，最终用户能够了解中央储存库加载期间，“他们”的数据发生了什么变化。

这种类型的工具通常能够很好地处理数据，所以我们很吃惊为什么这样的工具仍然内置在 IT 部门，而且一般不被数据挖掘者使用。*Mastering Data Mining* 一书中有一个 1998 年的案例，它使用 Ab Initio 的这类工具之一，分析上千亿字节的详细呼叫记录——即使在今天，处理如此海量的数据仍然是一项挑战。

15.2.3 中央储存库

中央储存库是数据仓库的中心，它通常是一个关系数据库，可以通过某些 SQL 的不同变体来访问数据。

关系数据库的优点之一是它们可以运行于功能强大、可以升级的计算机上，它们可以利用计算机的多处理器和磁盘阵列（参照后面“并行技术背景知识”部分）。例如，多数统计学软件包和数据挖掘包都能够同时进行多线程处理，每条线程代表一个任务，在一个处理器上运行。更多的硬件并不能让任何给定的任务运行更快（除非当其他任务恰好干扰它）。而关系数据库可以取单个查询，从本质上创建同时运行的多个线程来处理一个查询。最后的结果是，在强大的计算机上进行数据密集的应用时，使用关系数据库通常比使用非并行软件更快——数据挖掘正是一种数据密集的应用。

中央储存库的一个关键部分是逻辑数据模型，它以商业用户熟悉的术语描述数据库内部的数据结构。数据模型通常会与数据库的物理布局（或模式）相混淆，但两者之间有一个重要差别，物理布局的目的是为了使数据库的性能最优以及为数据库系统管理员（DBA）提供信息；而逻辑数据模型的目的是为了把数据库内容传达给更广泛、技术层次较低的受众。商业用户必须能理解逻辑数据模型——实体、属性及联系，物理布局是逻辑模型的一个执行工具，是沿着将性能最优化这一方向进行的折衷和选择。

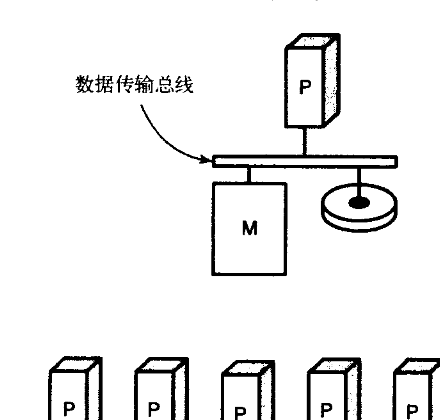
在从事一个数据仓库项目时，许多组织感觉迫切需要开发一个全面的、企业范围的数据模型。出乎意料的是，这些努力常常是失败的。数据仓库的逻辑数据模型不必像企业级模型一样不能作出某种妥协。例如，数据仓库的逻辑数据模型出现产品代码冲突，可以（但不一定必要）通过引入两种产品的分层结构来解决这个问题——仅仅花费 10 分钟就能进行的决定，而在企业级方案中，可能需要数月的研究与讨论。

提示：数据仓库是一个过程。要警惕任何一个作为数据仓库的大型数据库没有一个更新系统的过程来满足最终用户的需求。这样的数据仓库最终将被逐渐废弃，因为

最终用户的需求可能会逐渐发展，但这种系统却不能。

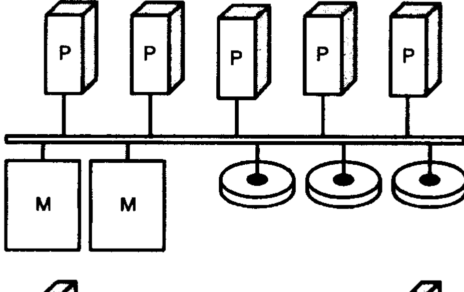
并行技术背景知识

并行技术是升级硬件的关键，它主要以两种情况出现：均衡多重处理系统（SMP）和大规模并行处理系统（MPP），这两种系统均被显示在下图中。SMP计算机以一条数据传输总线（bus）为中心（总线就是一个出现在所有计算机中的特殊网络，它将处理单元与内存和磁盘驱动器相连接）。数据传输总线作为中心通讯装置，因此SMP系统有时被称为完全共享系统，每个处理单元能够访问所有内存和所有磁盘驱动。这种并行方式是相当普遍的，因为SMP计算机支持像单一处理器计算机同样的应用——且某些应用可以利用额外的硬件而只对代码做最小的改变。但是，SMP技术有它的局限性，因为它给中心数据传输总线加上了一个很重的负担，当处理负载增加时它就会变得饱和。中心数据传输总线的争论点常常是限制SMP性能的那些内容。当处理单元少于10到20个时，它们趋向于工作得更好。



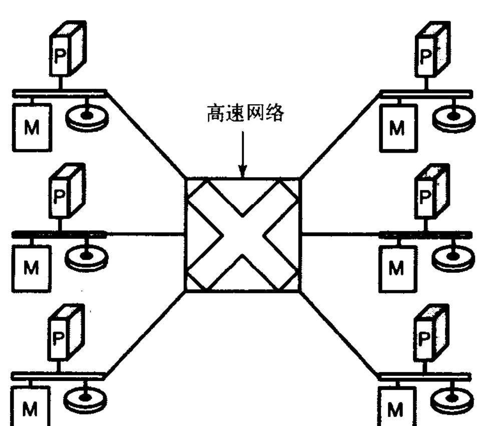
单处理器

一个简单的计算机遵循冯·诺依曼结构安排，处理单元通过一条局部数据传输总线与内存和磁盘通信（内存既存储数据又存储可执行程序）。处理器、数据总线及内存的速度限制了性能和可扩缩性。



SMP(symmetric multiprocessor，对称多处理器）

对称多处理器(SMP)有共享所有事情的结构，它扩展了数据传输总线支持多重处理器、更大内存及大型磁盘的能力，数据传输总线的这种能力限制了它的性能及扩缩能力。SMP结构通常利用最高不超过20个处理单元。



MPP(massively parallel processor，大规模并行处理器）

大规模并行处理器(MMP)有一个不共享的结构，它引入了高速网络（也称为交换机）将各自独立的处理器／内存／磁盘等组件连接起来。MPP结构非常容易扩缩，但仅有少数软件包可以充分利用所有硬件的性能。

并行计算机是以冯·诺依曼的单处理器结构为基础建立的。SMP和MPP系统是可扩缩的，因为更多的处理单元、磁盘驱动器、内存可以被增加到系统中

与之相对的是，MPP的行为像是把独立的计算机通过非常高速的网络（有时称为交换机）相连接。每一个处理单元有自己的内存及存储磁盘，某些结点可能是处理过程专用的，有最小的磁盘存储量；其他结点可能是为存储专用的，有非常大的磁盘容量。数据传输总线将处理过程单元与内存连接，而磁盘驱动器从来不会达到饱和。但存在的一个缺陷是，某些内存及磁盘驱动器是本机的，而另一些却是远程的——这个特点有可能使MPP很难进行编程。为一个处理器设计的程序总是可以在MPP中的一个处理器上运行——但它们要求进行某些修正以利用所有的硬件。只要连接处理器的网络可以提供更多的带宽，MPP的确可以升级，尤其是，更快的网络一般比更快的数据总线更容易设计。目前已经出现了有上千个结点和上千个磁盘的基于MPP的计算机。

SMP和MPP两者都有自己的优点。认识到这点以后，计算机生产商们尽量融合二者的优点。SMP生产商把SMP计算机联结在一起构成计算机群，开始模仿MPP计算机；同时，MPP生产商用SMP替代单一处理单元，产生非常相似的结构。然而，不管硬件是多么强有力的，软件仍需要进行优化设计以充分利用这些机器的性能。幸运的是，最大的数据库生产商已经投资数年，以使其产品能够满足要求。

数据仓库是用于管理记录的决策支持系统的一个过程。当用户需求随时间变清晰和发生变化时，这个过程能够按用户的需求来调整。用户的需求按时间变化时，这个过程能够对商业变化做出响应。如果没有意识到“当用户获知关于数据和商务的内容后，他们希望市场营销时间度量（数天和数周）也出现变化和增强，而不是增强IT的时间度量（数月）”这一点，中央储存库本身将是脆弱的、无用的系统。

15.2.4 元数据储存库

在前面有关数据层次的讨论中，我们已经讨论了元数据。它也可以被认为是数据仓库的组件。同样地，元数据储存库是一个常被忽视的数据仓库环境的一部分。元数据的最低层是数据库模式，即数据的物理布局。当正确使用的时候，元数据非常多。它回答了最终用户关于数据有效性的问题，为用户提供工具浏览数据仓库的内容，让每个人对数据更有信心，这种信心是新应用及扩大用户基的基础。

好的元数据系统将包括以下几个方面：

- 有注释的逻辑数据模型，这种注释应该说明实体和属性，包括有效值；
- 从逻辑数据模型到源系统的映射；
- 物理模式；
- 从逻辑模型到物理模式的映射；
- 访问数据的常用视图和规则，对一个用户有用的东西也许对其他用户也是有用的；
- 加载和更新信息；
- 安全和获取信息；
- 最终用户和开发者接口，以便共享数据库的相同描述。

在任何一个数据仓库环境中，这些信息片段中的每一个均可以在某些地方找到——在DBA写出的脚本中，在电子邮件、文件、数据库的系统表中等。元数据储存库可以让用户得到这些有用的信息，用一种他们容易理解的格式。关键就是给用户提供访问权限，以便他们能够方便地利用数据仓库，使用它所包含的数据以及知道如何使用它。

15.2.5 数据集市

数据仓库并非可以做任何事情（除了有效地存取数据以外）。应用必须实现价值，这通常以数据集市（data mart）的形式出现。数据集市是一个专用系统，它能够把部门或相关应用需要的数据结合在一起。数据集市常常被用于报表系统及分片切块数据。这样的数据集市经常使用 OLAP 技术，本章稍后将讨论这一问题。另一个重要的数据集市类型是一种用于数据挖掘的探测环境，这将在下一章中讨论。

并不是数据集市中所有的数据都需要来自中央储存库。通常，特殊应用对数据有独特的要求。例如，房地产部门可能正在把地理信息与中央储存库信息相结合；销售部门可能正把邮政编码人口统计学与中央储存库中的客户数据结合。中央储存库只需要包含不同应用之间可能共享的数据，因此它仅仅是一个数据源，对于数据集市来说，经常是占有主导地位的那个数据源。

15.2.6 操作反馈

操作反馈系统把由数据得来的决策返回到操作系统中。例如，一家大银行可能开发交叉销售模型以决定下一步提供给客户什么样的产品，这是数据挖掘系统的结果。然而，为了使它有用，这一信息需要返回操作系统。这就要求有一个从决策支持基础设施返回进入操作基础设施的联系。

操作反馈可提供快速完成有效数据挖掘循环的能力。一旦建立一个反馈系统，需要参与的工作仅仅是监测和改进它——为了让计算机做到最好（重复性的任务），让人们做到最好（发现重要的模式并提出想法）。以 Web 为基础的商业活动的优势之一是，从理论上讲，它们能够以一种完全自动化的方式为操作系统提供这样的反馈。

15.2.7 最终用户和桌面工具

在任何数据仓库中，最终用户是终极的和最重要的组成部分。没有用户的系统就没有创建的价值，这些最终用户是那些查找信息的分析师、应用软件开发者，以及依照信息进行商务活动的商业用户。

1. 分析师

分析师想访问尽可能多的数据来辨别模式和创建特定报告。他们使用专门的工具，像统计软件包、数据挖掘工具及电子数据表等。分析师常常被认为是数据仓库的主要受众。

通常，仅仅只有少数技术经验丰富的人属于这一类。尽管他们所做的工作很重要，但也很难判断一个基于增加生产力的大型投资是否正确。数据挖掘良性循环正是在这里开始起作用，数据仓库以干净的、有意义的格式把数据聚集在一起。尽管其目的是刺激创造力，但要测量它却是一个非常难以实现的想法。

分析师对数据仓库有非常特殊的要求：

- 系统必须能做出响应。众多的分析工作是以特定分析或特定疑问的形式来回答紧迫问题的方式进行的。
- 在整个数据库中数据必需相互一致。就是说，如果一个客户从某个特定日期开始，那么第一件产品、渠道等都应该准确地在那个日期出现。

- 数据必需在整个时期内一致。某个有特殊意义的字段在一定时间内回溯时必须有相同的意义。至少，其不同点应该有备份文件说明。
- 必须能够深入到客户层次、最好是交易层次的细节，来验证数据仓库中的值，并发展出关于客户行为的新的概要。

分析者给数据仓库加上了一个很重的负担，即必须用即时方式可以访问一致的信息。

2. 应用软件开发者

数据仓库通常支持一个宽阔的应用范围（换句话说，数据集市有各种各样的方式）。为了开发稳定及强有力的应用软件，开发者对数据仓库有一些特殊的要求。

首先，他们正在开发的应用软件需要与数据仓库的结构变化相隔离。新表、新字段及对现有表结构的改造，应该对现有应用的影响尽可能小。特殊的应用——特殊的视图帮助提供这项保证。另外，关于哪些应用使用哪种属性及实体的开放通信及知识能够防止出现拥塞僵局。

第二，开发者需要访问有效字段值，且需要知道该值代表什么意义，这是元数据库的目的所在，它提供数据结构中的文档说明。通过建立应用程序以元数据中的期望值校验数据值，开发者能够避免在应用软件完成后经常出现的问题。

开发者也需要在数据仓库的结构上提供反馈。通过识别必需包括在仓库中的新数据以及利用已经加载的数据来修正问题，是改进数据仓库的基本方法之一。因为真正的商务需要推动应用软件的发展，了解开发者的需求以确保数据仓库包含它所需要递送的商业价值的数据是重要的。

数据仓库将会发生变化，而应用软件将继续使用它。达到成功的关键是控制及管理这种改变。应用软件的目标是为了最终用户，数据仓库的目的是满足它们的数据需求——而不应是反过来。

3. 商业用户

商业用户是由公司数据仓库得来的信息的最终使用者。他们的需求推动着一系列方面的发展，包括应用软件、仓库的体系结构、所包含的数据及执行的优先权等。

多数商业用户仅通过印好的报告、静态的联机报告或电子数据表体验仓库——基本上与他们已长时间使用的收集信息的方式相同。即使如此，这些用户也会体验到拥有一个数据仓库的威力：报告变得更加精确、更加一致，而且更容易生成。

更重要的是，那些在办公桌上使用计算机的人们，乐意利用直接方式访问数据仓库环境。通常情况下，这些用户访问中间的数据集市来满足绝大多数信息需求，使用的是运行在他们熟悉的桌面环境上的友好图形工具。这些工具包括现成的查询生成器、客户应用软件、OLAP 界面及报告生成工具等。有时候，商业用户可以深入到中央储存库去探究那些在数据中发现的特别有趣的事情。更多的时候，他们会联系一个分析师，让他（或她）做一些更繁重的分析工作。

商业用户也会有针对特殊目的的应用软件，其中也许嵌入了前几章中讨论过的数据挖掘技术。比如，一个资源调度应用软件可能包含利用遗传算法来优化时间安排的引擎，一个销售预测应用软件可能含有内置的生存分析模型。当嵌入一个应用软件时，数据挖掘算法对最终用户通常完全隐藏起来，用户更加关心的是结果而不是产生结果的算法。

15.3 OLAP 适用于何处

几十年以来，商业领域一直在生成自动化的报告以满足商业需求。图 15-4 显示了各种常规报告手段。最古老的手工方法是大型机报告生成工具，它的输出传统上是打印在绿条纸或显示在绿色荧光屏上，这些大型机报告在计算机出现以前把基于纸的方法自动化。产生这样的报告通常是信息服务部门的基本职责，即使对报告进行很小的改变，也需要修改代码，有时甚至要耗用数十天。在用户要求变化时和用户看到几周及几个月后测量得到的新信息的时间之间，结果有一个延迟。这是很古老的技术，各个企业都试图避免这种情况，只有那些汇总特定操作系统的最低层的报告例外。

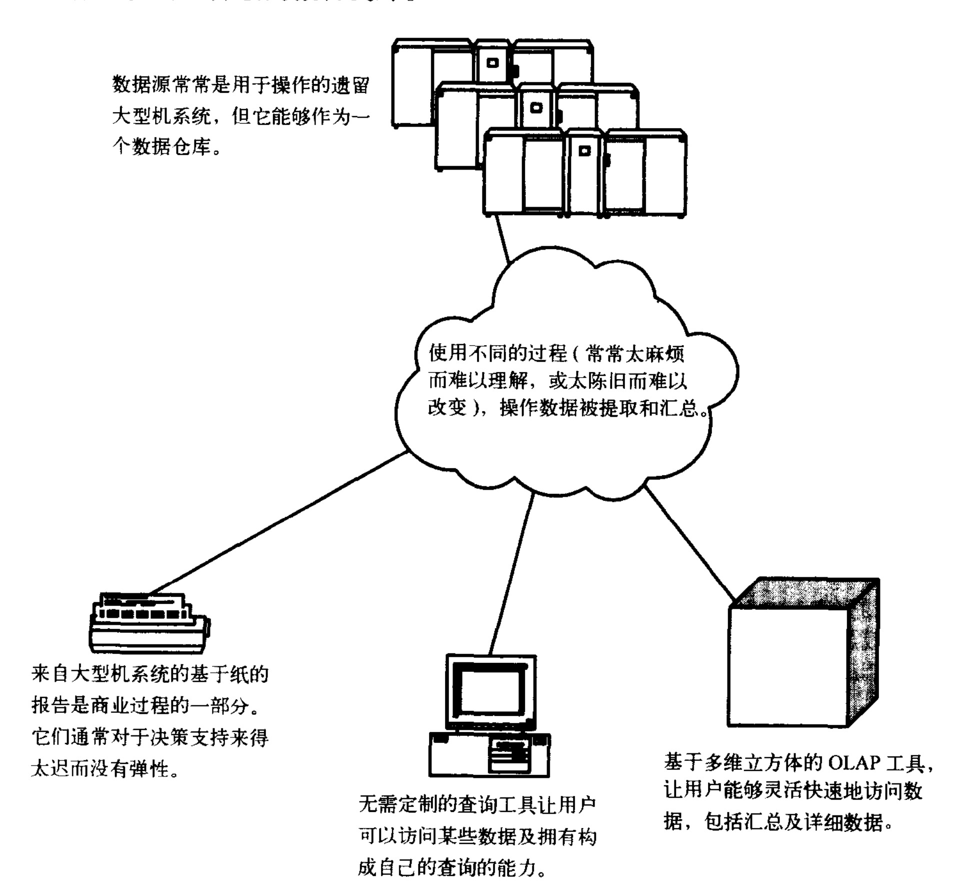


图 15-4 对操作系统的报告需求处理方式已经沿用了几十年，这是最佳方式吗

中间是不用定制的查询生成包，在过去的十年中，该查询包用于访问数据已经变得非常普遍。它们在 SQL 中产生查询，通过一个标准协议，比如开放式数据库连接（ODBC）标准，与局部的或远程的数据源进行对话。这种报告可能嵌入电子数据表中，通过 Web 进行访问，或通过某些其他的报告界面。利用大约一天时间的训练，商业分析师常常能够给出需要的报告。当然，报告本身常常作为一个 SQL 查询运行在已经超负荷的数据库中，所以当这种查询被认为运行完成时，响应时间通常是以分钟或小时为单位。这些响应时间比旧的报告生成包更快，但它们对于发掘数据仍然很困难。我们的目标是能够提出问题，而答案返回

时仍能记得该问题。

OLAP是一个针对特定查询系统的重要改进，因为OLAP系统设计数据结构时以用户为目的。这一强有力且高效的表示法被称为立方体，它非常适合分片及分块数据。该立方体本身或者存储在一个关系数据库中（典型的是星形模式），或者存储在一个由OLAP运算优化的多维数据库中。另外，OLAP工具提供了非常方便的分析功能，这在SQL中是很难或不可能进行的。如果说OLAP工具有一个不足的话，那就是它使得商业用户只关注于数据所代表的维。与之不同的是，数据挖掘对创造性思维是特别有价值的。

建立立方体要求对数据及最终用户需求进行分析，这一般要由熟悉数据和工具的专家通过一个称为多维建模的过程来完成。虽然设计和加载OLAP系统需要一项初始投资，但其结果为最终用户提供信息及快速访问，它通常比从查询生成工具得到的结果更加有用。一旦立方体建立起来，响应时间通常可以以秒计算，允许用户探究数据，刨根问底地理解他们所遇到的重要特征。

OLAP是对早期报告方法的强大改进，它的威力有三个关键特征：

- 第一，设计良好的OLAP系统有一组相关的维——比如地理、产品及时间等——这对于商业用户很容易理解。这些维对于数据挖掘目的通常是重要的。
- 第二，设计良好的OLAP系统有一组与商业相关的有用的度量。
- 第三，OLAP系统允许用户分片、切块数据，有时可下钻到顾客层次。

提示： *快速响应时间对获取用户对报告系统的认同非常重要。当用户需要等待时，他们可能忘记自己所问的问题，最终用户经历的交互响应时间应该在3～5秒。*

这些能力是对数据挖掘的补充，但不是它的替代。不过，OLAP是数据仓库结构中非常重要的（甚至可能是最重要的）部分，因为它拥有的用户数量是最大的。

15.3.1 立方体中的内容

了解OLAP的一个好方法是，把数据想象成把一个立方体分割成多个子立方体，如图15-5所示。虽然这个例子使用了三个维度，但OLAP可以有更多，三维对于说明目的是很有用的。这个例子显示了一个典型的零售业立方体，一维为时间，另一维为产品，第三维为店铺，每个子立方体包含各种度量，表示关于该种产品在某个日期正在发生的事情，如：

- 销售项目总数；
- 项目价值总和；
- 项目中折扣总和；
- 项目库存成本。

各种度量被称为事实。作为一项经验规则，维度是由分类变量（categorical variable）组成的，而事实是数值型的。当用户分片切块数据时，他们正从许多不同子立方体中聚集那些事实。维度被用于确定查询中使用了哪一个子立方体。

即使是上面所描述的一个简单的立方体，它也是非常有力的。图15-6给出了一个示例，它汇总了立方体中的数据，回答了“有多少天，一个特定商店没有销售一种特定产品”的问题。这样的问题需要使用商店及产品维度来确定哪个查询使用了哪个子立方体。这个问题仅考虑了一个事实，即售出项目的数目，返回所有该数据为0的那些日期。以下是另外一些可以相对容易回答的问题：

- 在过去的一年里，销售项目总和是多少？
- 以月为单位计数，东北地区的商店今年的销售与往年相比情况如何？
- 11 月每个商店的全部利润是什么（利润是指客户所付价钱减去库存成本）？

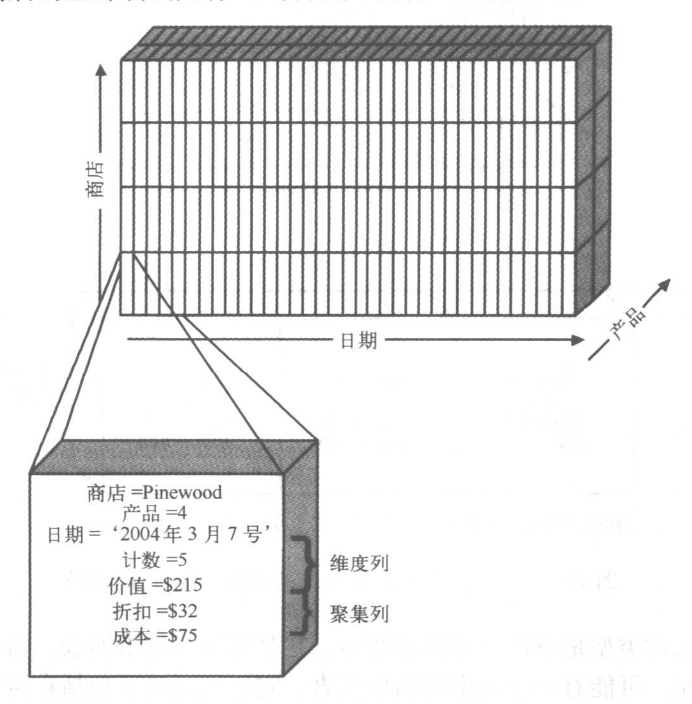


图 15-5 用于 OLAP 的立方体被分成多个子立方体，每个子立方体包含该立方体的键，以及落入那个子立方体的数据的概要信息

当然，获取能够回答这些问题中的一个问题的报告的难易程度，取决于该报告界面的具体执行方式。然而，即使对于特殊报告，访问立方体结构比访问一个规范化关系数据库来说要更容易一些。

1. 立方体的三种变体

上一节所描述的立方体是一个概要数据立方体的例子。这是在 OLAP 中很常见的例子。然而，并不是所有的立方体都是概要立方体，而且，针对不同目的，一个数据仓库可能包含许多不同的立方体。

另外一种类型的立方体代表的是个别事件。这些立方体包含着与客户互动的最详细的相关数据，如呼叫客户服务、支付、个人账单等。概要可以通过聚集整个立方体的事件得到，这种事件立方体通常有一个客户维度或某些相似的事情，如账号、Web cookie 或一个家庭，这些事情可以把事件与客户联系起来。通常，很少数目的维度，如客户 ID、日期、事件类型就足以识别每个子立方体。然而，一个事件立方体常常还有其他一些维度，它们提供更详细的信息，对于聚集数据很重要。在这样的一个表中的事实，常常包含美元总数和计数。

事件立方体功能是非常强大的，但它们的应用受限，因为它们能飞快地变大——代表它们的数据库表可能有数百万、数亿甚至数十亿行，即使利用 OLAP 及并行计算机的计算能力，这样的立方体对于常规查询还是需要一些处理时间。虽然如此，事件立方体仍有特殊的价值，因为它们使得从其他立方体“下钻”进去成为可能——去发现精确的事件集用于计算特定值。

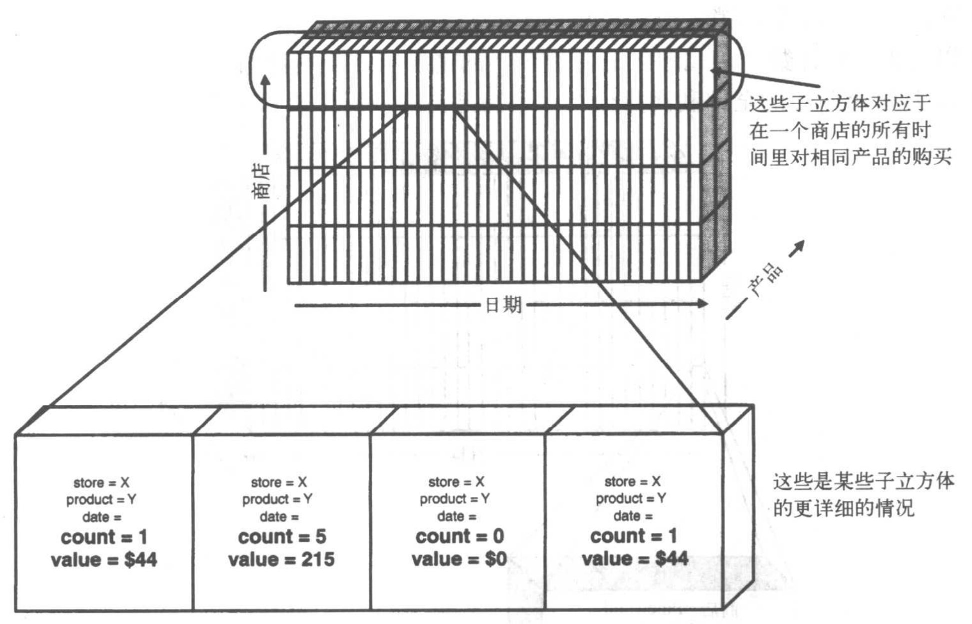


图 15-6 在过去的多少天中 X 商店没卖出任何产品 Y

立方体的第三种类型是事件立方体的变体。它是无事实的事实表，目的是代表某些事件发生的证据。例如，可能有一个无事实的事实表，它详细说明了包括在直接邮寄活动中的潜在顾客，这样的事实表可能有下列一些维数：

- 潜在客户 ID（或许是家庭 ID）；
- 潜在客户来源；
- 邮寄的目标日期；
- 消息类型；
- 创造性类型；
- 优惠类型。

这就是对于某个姓名可能不存在任何数值型事实的情况。当然，对于维度，可能有令人感兴趣的特征——例如该优惠的促销花费和购买名单的花费等。但这个数据可以通过维度来得到，因而不需要在单个潜在客户的层次上进行重复。

不管事实表的类型如何，都有一个重要的规则：任何一条特别信息都应该刚好落入一个子立方体中。违反了这一规则，立方体就不能被容易地用于各种各样维度的报告。这一规则的一个必然结果是：当一个 OLAP 立方体被加载，追踪任何出现意外维度值的数据是非常重要的。每一个维度应该有一个“其他”类，以保证所有进来的数据有位置。

提示：当为立方体选择维时，要确信每一条记录处在一个准确的子立方体中。如果有多余的维——如一个维是日期，另一个维是周几——那么同一记录将处在两个或更多的子立方体中。如果发生了这样的事，那么基于子立方体的汇总就不准确。

除了插入立方体的每条记录应该刚好处于一个子立方体中这条最为重要的基本规则以外，设计有效的立方体时，要谨记以下另外三件事情：

- 确定事实；
- 处理复杂维度；
- 使维度在整个数据仓库中保持一致。

当试图发展立方体的时候，就会出现这三个问题，解决它们对于立方体用于分析目的是很重要。

2. **事实**

事实就是每一个子立方体的度量。最有用的事实是可以求和的，因而它们可以把许多不同的子立方体结合在一起，从而在任何汇总层次上提供查询的响应。可求和的事实可以使我们在任意维度方向上或同时沿几个不同的维度来汇总数据——这恰恰是使用立方体的目的之所在。

可求和事实示例：

- 计数；
- 具有一个特定值的变量计数；
- 合计持续时间（如花费在某个 Web 站点上）；
- 合计币值。

某一天花费在某件特定商品上的总金额就是每一家商店花费在该产品上的金额总数，这是一个可求和事实的好例子。但不是所有事实都是可求和的，不可求和的例子包括：

- 平均值；
- 惟一计数；
- 不同立方体共享事物的计数，如交易。

平均值并不是一个不可求和事实的重要的例子，因为平均值是总数除以计数。由于其中的每一个都是可求和的，可以结合这些事实以后来导出平均值。

其他实例更有意义，一个重要的问题是有多少独特的客户做某个特别的举动。虽然这些数值能够在子立方体中存储，但不是可求和的。考虑具有日期、商店、产品维度的一个零售立方体。个别客户可能在多个商店中购买项目，或在一个商店购买多个项目，或在不同的时间进行购买。包含独特客户数目的字段有关于某个客户在多个子立方体中的信息，这违反了 OLAP 最重要的规则，因此，该立方体将不能报告独特的客户。

当试图计算交易数目时会发生类似的事情。因为关于交易的信息可能存储在几个不同的子立方体中（因为个别交易可能包括多项产品），交易计数也违反了该重要规则。这种类型的信息不能在概要层上收集。

关于事实需要注意的另一个问题是，并非所有数值型数据都可作为立方体的事实。例如，多少岁是数值型的，但把它作为一个维比作为一个事实更好。另一个例子是客户价值，把客户价值的离散范围当成一个维更有用，在许多情况下会比试图将客户价值当成事实更有用。

当设计立方体的时候，为一组相关的数值创建一个计数或总和，很容易把事实与维度混淆到一起，例如：

- 保有期少于 1 年、在 1 到 2 年之间、超过 2 年的活跃客户计数；
- 在每周工作日中的花销数量，在每个周末的花销数量；
- 某周中每天的总数。

上述中的每一个都建议立方体建立另外的维，第一个应该有一个客户保有期维，它至少有三个值；第二个出现在一个以月作为时间维的立方体中，这些事实建议需要一个每日概要，至

少可以沿着一个维分离工作日和周末；第三个建议需要一个以天数间隔为单位的日期维。

3. **维及其分层**

有时，单个列似乎适于多个维。例如，OLAP 对于可视化按时间变化的趋势是一个好的工具，像销售数据或财务数据等。在这个示例中，一个特定日期潜在地表现了沿几维的信息，如图 15-7 所示：

- 周几
- 月份
- 季度
- 历法年

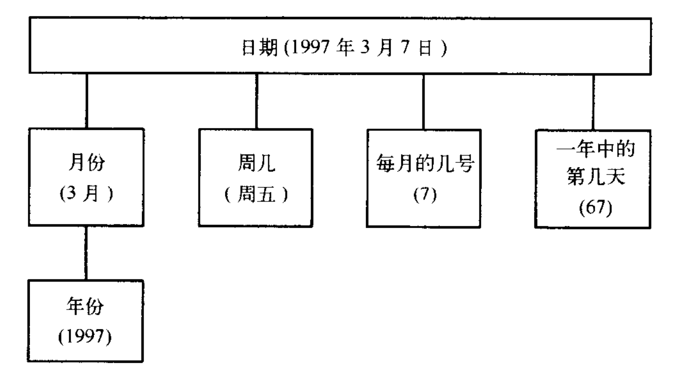


图 15-7 日期的多重分层

一种方法是把每一个都作为一个不同的维。换句话说，可能有四个维度，一个是周几，一个是月，一个是季，还有一个是历法年。这样，2004 年一月的日期就会是一个子立方体，即一月份维度与 2004 年维度相交的地方。

这不是一个好方法。多维建模发现，时间是一个重要的维，并且可以有很多不同的属性。除了以上所述的属性以外，还有该周属于一年中的哪一周，该日期是否是假日，该日期是否是工作日等。这样的属性存储在参照表中，称之为维表。维表使我们可以改变维的属性而不改变根本的数据。

警告：当为一个 OLAP 系统设计维度的时候不要找捷径，确实存在一些数据集市的框架，一个不牢靠的框架不会持续很长时间。

维表包含许多不同的属性，描述该维的每个值。例如，一个详细的地理维可能由邮政编码创建，它包括关于邮政编码的几十个概要变量。这些属性能够用于过滤（“有多少客户在高收入地区?”）。这些值被存储在维表而不是事实表中，因为它们不能正确地聚集。如果在一个邮政编码区域中有三个商店，一个邮政编码的人口事实将会求和三次——总人口乘以 3。

通常，维表以维的最新数值更新。这样，商店维可能包括当前的一组商店以及关于商店的信息，如布局、面积大小、地址、经理姓名等。然而，所有这些可能按时间发生变化。这样的维被称为缓慢变化维，它对于数据挖掘有特殊的意义，因为数据挖掘意图重建精确的历史。缓慢变化维超出了本书的范围，有兴趣的读者可以翻阅 Ralph Kimball 的书。

4. **一致维**

正如前面提到的，数据仓库系统通常包含多重 OLAP 立方体。OLAP 的某些功能就是从

共享不同立方体维的实践中发现的。这些共享维称为一致维，如图 15-8 所示。它们有助于确保通过不同系统报告的商业结果应用一组相同的基本商业规则。

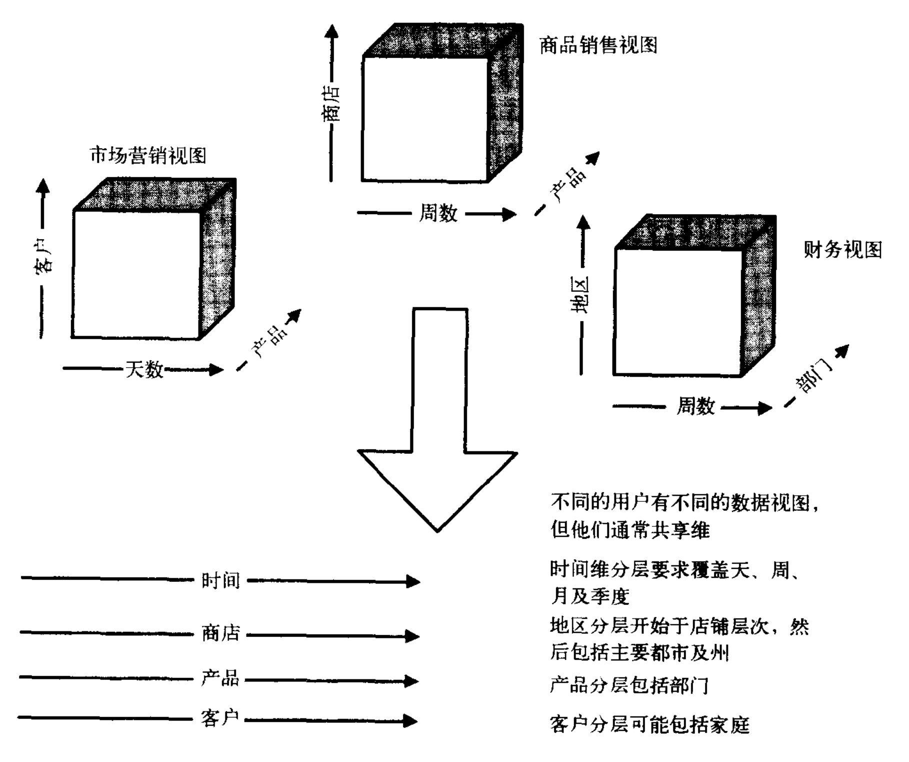


图 15-8 对数据的不同视图通常共享共同的维，发现共同的维及其基本单元对于使数据仓库在整个组织良好地运转是非常重要的

一致维的一个典型例子是日历维，它追踪每天的属性。日历维非常重要，它应该是每个数据仓库的一部分。然而，数据仓库的不同成份可能要求不同的特征，例如，一个跨国公司可能包括不同国家的几套节假日，因此应该有"美国节假日"、"大不列颠联合王国节假日"、"法国节假日"等标记，而不是总体的假日标记。在许多国家 1 月 1 日是节日，然而，7 月 4 日在美国是最重要的节日。

创建 OLAP 系统面临的挑战之一是设计一致维，以便它们适用于一系列广泛的应用。针对某些目的，地理位置可能用城市和州来描述最好；而有些则可以用国家来描述；另一些用人口普查区域描述；还有一些通过邮政编码来描述。不幸的是，这四种描述方法不完全一致，因为在邮政编码中可能有几个小镇，在纽约市有 5 个郡。多维建模可以帮助解决这种冲突。

15.3.2 星形模式

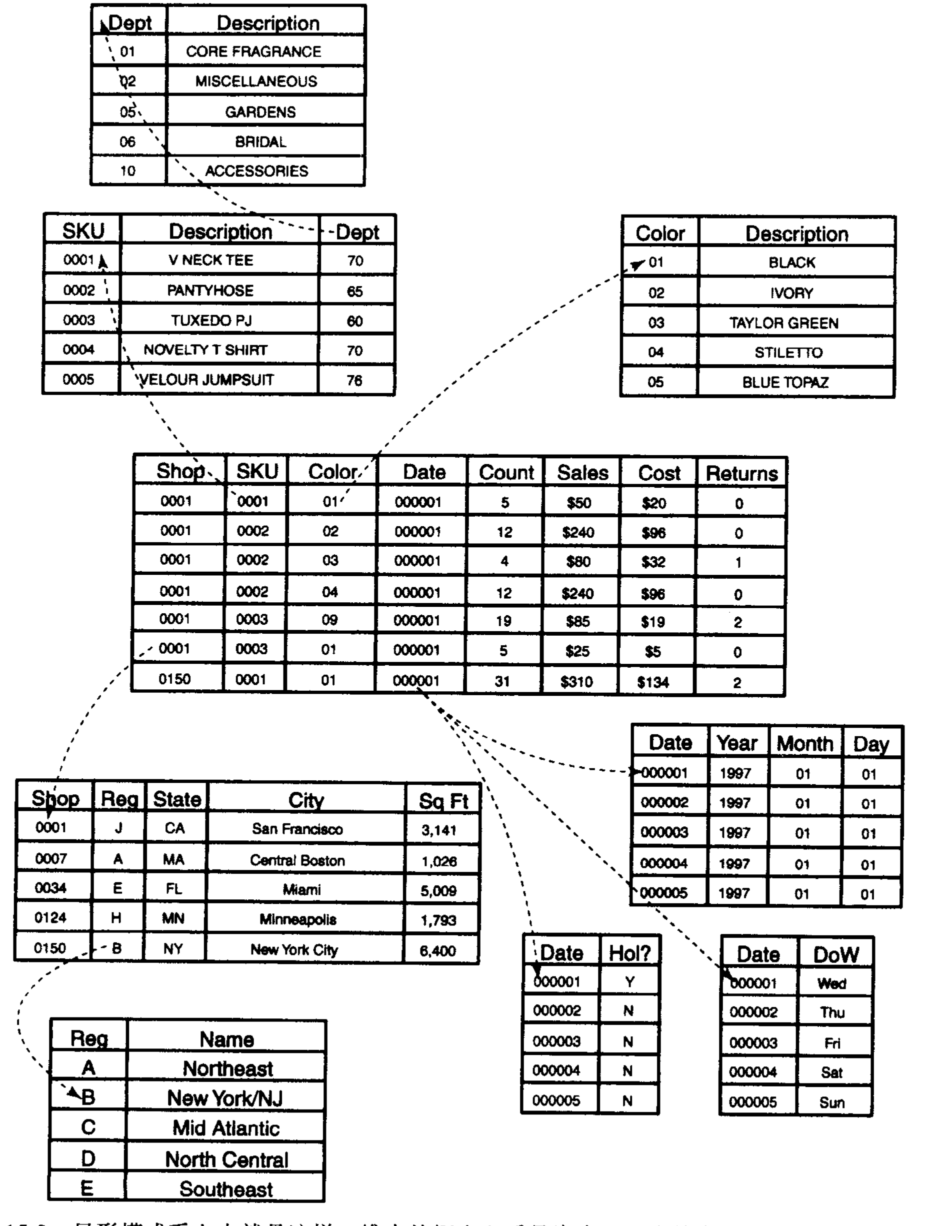
使用被称为*星形模式*的非标准化数据结构，很容易在关系数据库中存储立方体，这是由 OLAP 的一位宗师 Ralph Kimball 设计的。星形模式的一个优点是它可以使用标准数据库技术实现 OLAP 的强大功能。

星形模式始于一个对应于商业事实的*中心事实表*。这些可能处于交易层次（对于一个事件立方体），尽管它们常常是较低层次的交易概要。对于零售业务，中心事实表可能包含每种产品在每个商店的日销售（shop-SKU-time）概要。对于一个信用卡公司，事实表包含的

行对应于每个客户的每一笔交易，或者是基于产品（基于卡的类型及信用限制）、客户片段、商业类型、客户地理位置及月份的花费概要。对于一个对修理历史感兴趣的柴油机制造商，它可能包含对每台机器的每次修理，或在每个商店按照修理类型给出的每日修理概要。

在中心事实表中的每一行包含使它惟一的一些键的组合。这些键称为维（dimension）。中心事实表也有其他列，通常包含对应于每一行的数值型信息，如交易总量、交易数目等。与每一个维关联的辅助表格称为维表（dimension table），它包含对应于维的特别信息。例如，日期的维表可能详细说明某个特定的日期是周几、月份、年份以及是否节假日。

在图表中，维表与中心事实表相连，结果在形状上很像一个星星，如图 15-9 所示。



Dept	Description
01	CORE FRAGRANCE
02	MISCELLANEOUS
05	GARDENS
06	BRIDAL
10	ACCESSORIES

SKU	Description	Dept
0001	V NECK TEE	70
0002	PANTYHOSE	65
0003	TUXEDO PJ	60
0004	NOVELTY T SHIRT	70
0005	VELOUR JUMPSUIT	76

Color	Description
01	BLACK
02	IVORY
03	TAYLOR GREEN
04	STILETTO
05	BLUE TOPAZ

Shop	SKU	Color	Date	Count	Sales	Cost	Returns
0001	0001	01	000001	5	$50	$20	0
0001	0002	02	000001	12	$240	$96	0
0001	0002	03	000001	4	$80	$32	1
0001	0002	04	000001	12	$240	$96	0
0001	0003	09	000001	19	$85	$19	2
0001	0003	01	000001	5	$25	$5	0
0150	0001	01	000001	31	$310	$134	2

Date	Year	Month	Day
000001	1997	01	01
000002	1997	01	01
000003	1997	01	01
000004	1997	01	01
000005	1997	01	01

Shop	Reg	State	City	Sq Ft
0001	J	CA	San Francisco	3,141
0007	A	MA	Central Boston	1,026
0034	E	FL	Miami	5,009
0124	H	MN	Minneapolis	1,793
0150	B	NY	New York City	6,400

Date	Hol?
000001	Y
000002	N
000003	N
000004	N
000005	N

Date	DoW
000001	Wed
000002	Thu
000003	Fri
000004	Sat
000005	Sun

Reg	Name
A	Northeast
B	New York/NJ
C	Mid Atlantic
D	North Central
E	Southeast

图 15-9 星形模式看上去就是这样。维表从概念上看是嵌套，一个给定的维可能有多个维表

事实上，星形模式不可能有效地回答所有用户的问题，因为中心事实表太大。在这种情况下，OLAP 系统在不同层上引入了概要表以方便查询响应。关系数据库生产商已经对星形模式提供了越来越多的支持。利用一个典型的结构，对中心事实表的任何查询将需要多次连接回维表。通过应用标准索引、创造性的增强索引技术，关系数据库能够相当好地处理这些查询。

15.3.3 OLAP 和数据挖掘

数据挖掘是为决策支持目的而对数据进行的成功开发。在第 2 章中描述的关于数据挖掘的良性循环，提醒我们成功更多地依赖于先进的模式识别算法。数据挖掘进程（data mining process）需要给人们提供反馈，鼓励使用从数据挖掘中得到的信息以改善商业过程。数据挖掘进程可以使人们提供输入，以观测和假设的形式，预测什么结果是重要的及如何应用那些结果。

从数据开发的更大范围来看，OLAP 作为一种拓宽数据访问受众的手段，很明显扮演了一个重要的角色。以前基于经验和猜测制定的决策，现在可以基于数据和数据中的模式。异常点和离群值（outlier）可以被识别出来以用于深入研究和进一步建模，有时候需要使用最复杂的数据挖掘技术。例如，通过使用 OLAP 工具，某个用户可能发现在某周的一个特定时间某个特定项目销售得更好。这可能导致使用购物篮分析发现其他可能与该物品同时购买的物品的研究，购物篮分析可能对于观察到的行为给出一个解释——更多信息和利用这些信息的更多机会。

数据挖掘和 OLAP 之间有其他方面的协同作用。在第 6 章中讨论的决策树的特点之一是，识别在数据中与特殊结果有关的最具信息价值特征的能力。就是说，如果一个决策树是为了预测流失而创建，那么树的上层将会有最重要流失预测器的特征。这些预测器可能是使用 OLAP 工具时维的一个好选择。这种分析可以帮助建立更好、更有用的立方体。建立立方体时，另外的问题是确定如何使连续的维离散。决策树的结点能够帮助确定一个连续值的最好断点。这一信息能够馈入 OLAP 工具以改善维。

神经网络遇到的问题之一是难于理解结果。当把它们用于非定向数据挖掘时更是如此，就像使用 SOM 网络来探测聚类那样。SOM 可以识别聚类，但不能解释聚类的意义。

这时，OLAP 就可以过来救援！数据现在能够利用预测的聚类增强，也可利用关于聚类的其他信息，如人口统计学、购买历史等。这对于立方体是一个很好的应用。使用 OLAP 把关于簇的信息作为一个维，使最终用户可以探测簇，确定辨别它们的特征。用于 OLAP 立方体的维应该包括进入 SOM 神经网络的输入，还有簇标识符以及其他可能的描述性变量。这里存在一个数据转化的技巧问题，因为神经网络需要缩放到 -1 和 1 之间的连续值，而 OLAP 工具则更喜欢离散值。对于初始离散数据，这没有问题；但对于连续数值，各种分箱（binning）技术解决了这个问题。

就像这些例子中显示的那样，OLAP 和数据挖掘相互补充，通过定义合适的维（更进一步，通过确定在维中如何断开连续值），数据挖掘能够帮助建立更好的立方体。OLAP 提供了一个强大的可视化能力，以帮助用户更好地理解数据挖掘结果，如聚类和神经网络。联合使用 OLAP 和数据挖掘，二者彼此补充了它们的实力，为数据开发提供了更多的机会。

15.4 数据挖掘在哪里切入数据仓库

数据挖掘在数据仓库环境中扮演着很重要的角色。数据仓库的初始回报来自已有的自动化过程，如使报告在线，给已有的应用软件一组清洁的数据源等。最大的回报是改善数据的访问，这能够激励改革和创新——这些都源于观察和分析数据的新方式。这就是数据挖掘所扮演的角色——提供工具以改善理解，并基于对数据的观察激励创新。

一个好的数据仓库环境可以作为数据挖掘的良好催化剂，这两项技术可以一起使用：

- 数据挖掘在大量数据基础上茁壮成长，数据越详细越好——而数据来自数据仓库。
- 数据挖掘在清洁和一致的数据中茁壮成长——利用了数据清理工具。
- 数据仓库环境支持假设测试，简化了测量所采取行动的效果的工作——支持数据挖掘的良性循环。
- 可升级硬件和关系数据库软件能够分担数据挖掘的数据处理部分。

然而，数据挖掘的观点与数据仓库的观点存在差别。规范化的数据仓库能够引入时间戳来存储数据，但是进行时间相关的处理是非常困难的——比如确定什么事件正好会在其他感兴趣的事件之前发生。OLAP 引入了一个时间轴，数据挖掘把这种思想扩展得更远，甚至可以考虑“之前”和“之后”的概念。数据挖掘从数据（即“之前”）中学习，目的是把学到的知识用于未来（即“之后”）。正是由于这个原因，数据挖掘经常给数据仓库带来巨大的工作量。它们属于互补的技术，正如稍后的几小节会讲到的，它们互相支持。

15.4.1 大量数据

数据分析的传统方法一般从减小数据量的规模开始，通常有三种方法：汇总详细的交易数据，从数据中取出一个子集，以及只观察某些属性。减小数据量大小的原因在于，我们可以在现有的软硬件系统上分析数据。当这些问题得到合理处理以后，统计学的定律就可以引入，从而有可能选出一个行为表现大致接近于其他数据的样本。

另一方面，数据挖掘寻找数据的趋势，寻找有价值的异常点。它常常试图回答传统的统计分析提出的不同类型的问题，如“什么样的产品是这一客户下次最可能购买的?”即使可能使用一个数据子集设计模型，也必须配置该模型，为所有客户打分，这是一个计算量非常大的过程。

幸运的是，数据挖掘算法常常能够利用大量的数据。当寻找模式以识别稀有事件时——如不得不勾销客户，因为他们没有付款——有大量数据就保证有足够的数据进行分析。一个数据的子集从统计学角度看总体上可能是恰当的，但当你试图把它分解为其他片段（按地区、产品、客户片段）时，要给出有统计意义的结果，数据就有可能太少。

数据挖掘算法能够利用大量数据。例如决策树，即使当每一条记录中有数十或数百个字段时，它也可以很好地工作。链接分析要求用一个完整的数据集创建图。神经网络能够在同一时间训练数百万记录。这些算法常常运用详细交易的汇总来工作（特别是在客户层次上），汇总结果可能由这次运行到下次运行发生改变。预先创建汇总和丢弃交易数据将把你锁定到一个商业视图上。当然，应用这种汇总的第一个结果经常导致需要对它们进行某些变化。

15.4.2 一致的、清洁的数据

数据挖掘算法常常需要用到吉字节（1 吉字节 = 10^9 字节）的数据，这些数据可能来自

几个不同的源。在寻找可操作的信息时，大部分工作实际上是把数据结合在一起——通常数据挖掘项目的时间有80%或更多是用于把数据汇集到一起——特别是当没有数据仓库可用的时候。此后的问题，如匹配账号、翻译代码、分拆等，会进一步拖延分析。发现重要的模式通常是一个交互过程，需要返回到数据以获得另外的数据元素。最终，当发现重要模式的时候，通常需要在最近可用的数据上重复这个过程。

一个设计良好及构造良好的数据仓库能够帮助解决这些问题。当数据被加载到数据仓库时，数据被清理一次。字段的意义被明确定义且可以通过元数据利用。将新数据整合到分析中就像通过元数据找出什么数据可用以及从数据仓库重新找回一样容易。一个特别的分析能够重新运用于更近的数据，因为数据仓库一直保持最新。最终结果是数据更清洁、更好用——这使分析师可以花更多的时间应用功能强大的工具和洞察力而不是移走数据和压缩数据量。

15.4.3 假设测试和测量

数据仓库推动了数据挖掘的两个其他领域。假设测试是验证数据中关于数据模式的基于经验的猜测。热带色彩在佛罗里达确实比在别处更好销售吗？人们倾向于在晚饭后打长途电话吗？在餐馆的信用卡用户确实是高端客户吗？所有这些问题都可以非常容易地在适当的关系数据库中作为查询表达出来。拥有可利用的数据使提问及快速发现答案成为可能。

提示：测试假设和思想的能力是数据挖掘一个非常重要的方面。通过将数据结合在一起，数据仓库能够深入地回答复杂的问题。一个要注意的问题是这样的查询运行代价昂贵，从而陷入杀手查询类。

测量是另一个已证明数据仓库非常有价值的应用领域。通常，当进行市场营销、产品改进等工作的时候，达到成功的程度只有一个有限的反馈。数据仓库可以看到结果并发现相关的影响。其他产品销售得到改进了吗？客户流失是否有所增加？打到客户服务中心的电话减少了吗？等等。有可用的数据使得理解一个行动的结果成为可能，不管行动是通过数据挖掘结果激励的还是其他事情激励的。

从测量的角度说，特别有价值的是不同市场营销行为对中长期客户关系的影响。通常，市场营销活动是以响应率来测量的。然而响应率只是人们感兴趣的一个方面，仅仅是其中一个。客户的中长期行为也是令人感兴趣的内容之一。获取活动带来好的客户了吗？或者新获得的客户在没付款之前已经离开了？提升销售活动坚持住了吗？或者客户又回到了老产品？测量可以使某个企业从它的失误中吸取教训，并走向成功。

15.4.4 可升级硬件及RDBMS支持

数据挖掘和数据仓库之间的最终协作是在系统层次上，同样的可升级硬件和软件使得储存和查询大型数据库成为可能，这为分析数据提供了一个好的系统。第17章讨论创建客户特征标识，通常创建特征的最好地方是在中央储存库，或者，如果没有的话，可在数据量相似的数据集市中。

进一步利用功能强大的计算机的优势，并行运行数据挖掘算法仍然是一个问题。这通常没有必要，因为实际上建立模型只代表了数据挖掘时间投入的一小部分——准备数据及理解结果更加重要。一些数据库，像Oracle和微软的SQL Server，正不断为数据挖掘算法提供支

持，这使得这样的算法能够并行地运行。

15.5 小结

数据仓库不是一个系统，而是一个对于数据挖掘和数据分析工作非常有用的方法。从数据挖掘的观点看，最重要的功能是再创造历史的准确快照的能力。另一个非常重要的方面是支持特定报告。为了从数据中学习，你需要知道究竟发生了什么。

典型的数据仓库系统包含下列部分：

- 源系统提供到数据仓库的输入；
- 提取、转化和加载工具清理数据并应用商业规则，以便新数据与历史数据相一致；
- 中央储存库是一个特别为记录的决策支持系统设计的关系数据库；
- 数据集市为具有不同需求的不同用户提供界面；
- 元数据储存库告知用户和开发人员数据仓库内是什么。

数据仓库面临的挑战之一是必须储存大量的数据，特别是当目标是保持所有客户的交互数据的时候。幸运的是，计算机有这个能力，问题只是更多的预算而不是可能性。关系数据库也能够利用最强有力的硬件——并行计算机。

联机分析处理（OLAP）是数据仓库的一个强有力的部分。OLAP 工具擅长处理汇总数据，允许用户一次沿一维或几维汇总信息。因为这些系统是为用户报表优化而设计的，通常它们的交互响应时间少于 5 秒。

任何设计良好的 OLAP 系统有一个时间维，这使得它观察按时间变化的趋势时非常有用。而在一个规范化的数据仓库中完成同样的一件事情需要非常复杂的查询，而且容易出错。更有用的是，OLAP 系统将允许用户对所有报告深入到详细数据中。这种能力确保所有数据进入立方体中，也给用户提供了发现可能没有出现在维中的重要模式的能力。

就像我们贯穿本章所指出的那样，OLAP 补充了数据挖掘，但不是数据挖掘的替代。它提供了对数据的更好理解，且为 OLAP 开发的维能够使数据挖掘结果更加具有可操作性。然而，OLAP 不能在数据中自动发现模式。

OLAP 是向最终用户提供高级报告需求的一种强有力的信息发布方式。它提供了这样的能力，即让更多的用户基于数据进行决策，而不是依靠预感、猜测和个人经验。OLAP 补充了像聚类这样的非定向数据挖掘技术。OLAP 能够提供在识别出的簇中找到商业价值所需要的洞察力。它也提供一个好的可视化工具，可用于其他方法，如决策树及基于存储的推理。

数据仓库与数据挖掘不是一回事，然而，它们相互补充，数据挖掘应用往往是数据仓库解决方案的一部分。

第16章 构造数据挖掘环境

在一座大块冰糖山上，
有一个美丽而快乐的桃花源，
灌木丛中生长着衣服、食物和金银。
人们喜欢每晚露天而宿，
所有房车都是空的，
每天阳光普照。
鸟儿和蜜蜂快乐飞翔，
香烟满树，
柠檬汁成泉，
知更鸟歌唱，
在一座大块冰糖山上。

20世纪的流浪者有这样的乌托邦式幻想，21世纪的数据挖掘者为什么不也幻想一下呢？对我们来说，这种幻想就是一家公司将客户放在运作的中心，并通过长期的客户价值结果测量其行为。在这一完美的组织中，商业决策是基于从大量客户数据中提取的可靠信息。不必说，数据挖掘者——拥有将所有数据转化为公司运行所需要信息的技巧的人们——会赢得最高的尊重。

本章从一个真正的以客户为中心的（customer-centric）乌托邦式幻想组织开始，在那里有理想的数据挖掘环境，产生的信息能够作为决策的基础。了解一个理想的数据挖掘环境是什么样的，有益于建立更真实的近期目标。然后，本章继续在数据挖掘环境的各种组成中寻找——人员、数据挖掘的基础设施、数据挖掘软件本身。尽管不可能达到乌托邦幻想所有的元素，但可以利用这种幻想帮助创建一个适合数据挖掘工作的环境。

16.1 以客户为中心的组织

尽管大家都说客户就是上帝，但在大多数公司中并没有把客户当成上帝。一个原因是大多数商业不是围绕客户组织，而是围绕产品组织。例如，超市长期以来能够追踪成千上万种产品的详细库存信息，以便保持货架供应，并且能够在任何项目上计算利润额。但直到最近，这些商店对于每位客户却一无所知——不知道他们的姓名，也不知道每月他们来几次，他们倾向于在每天的何时购物，是否使用商家的优惠券，是否有孩子，在这个商店中某个家庭购货的百分比是多少，他们住得多近——什么也不知道。我们无意对超市进行挑剔，银行围绕借贷组织，电话公司围绕交换机组织，航班围绕运转组织，没有人知道（或关心）客户太多。

在所有这些行业，技术的发展使他们可能把焦点转移到客户。这样的转移并不容易，事实上，这完全是一场革命。通过将这些销售点（point-of-sale）扫描器的数据与忠诚卡计划相结合，杂货零售商通过一些努力，能够知道谁购买了什么及他们何时购买的，哪些客户对价格敏感，哪些喜欢尝试新产品，哪些喜欢自制面点，哪些更喜欢半成品等；电话公司能够断

定谁在进行商业呼叫，谁主要与朋友聊天；在线音乐商店能够向顾客推荐个性化的新音乐。

更艰巨的挑战是能够有效地利用这项能力观察数据中的客户。一个真正以客户为中心的组织会很愿意继续提供无利可图的服务——如果使用这种无利润服务的客户在其他领域肯花费更多，就可以在总体上增加公司的收益。一个以客户为中心的公司不必每次当客户呼入时都问同样的问题。以客户为中心的公司以顾客在整个生存周期中产生的价值来评估市场营销活动，而不是初始响应率。

要做到真正的以客户为中心意味着要改变公司文化，改变从上层管理者到客服中心接线员等每个人的奖励方式。只要每一条生产线有一个管理者的薪水与产品销售的数量及利润挂钩，该公司将仍然把注意力集中在产品上而不是客户上。换句话说，公司雇佣管理人员关注产品，而管理员做了他们该作的工作。在一个理想的以客户为中心的组织中，每个人都因为增加客户价值而受到奖励，他们知道这需要通过与每个客户进行交互，而且应该具有用所获得的知识更好地为客户服务的能力。结果，公司记录了与客户的每一次交互，并保留了这些交互广泛的历史记录。

16.2 理想的数据挖掘环境

理想的数据挖掘环境是一个能够正确评价信息价值的组织。把从许多收集原始数据的地方得到的客户数据汇集在一起，并把它们处理成适合数据挖掘的形式是一个困难且耗资巨大的过程。它只可能发生在知道数据一旦使用得当会多么有价值的一个组织中。信息就是力量。一个会学习的组织崇尚的是进步和稳定的改善，这样的组织希望而且会为准确的信息投资。记住，信息的制造者常常有真正的权力去决策什么样的数据在什么时候是可用的，他们不是只决定取舍的数据仓库的被动消费者，他们有能力决定什么数据是可用的，虽然收集这样的数据可能意味着要改变操作程序。

16.2.1 确定什么数据可用的能力

在理想的数据挖掘环境中，数据分析的重要性是被认可的，数据分析的结果在整个组织中是共享的。从事市场营销的人们从本能上把每一项营销活动看做一个对照实验，这甚至意味着在一个有良好预期效应的营销活动中不包括某些客户，因为那些客户是一个对照群组(control group)的一部分。操作系统的设计者本能地追踪所有客户事务，包括那些不需账户支付的事务，如客户服务查询、银行账号结算查询或访问公司Web站点的特殊部分。每一个人都期望当涉及同一位客户时，不同渠道的客户交互能够被识别，即使某些交互发生在ATM，某些发生在银行分行，某些通过电话，某些是通过Web。

在这样的环境中，电话公司的一位分析员在试图了解无线电话服务质量及客户流失之间的关系时，将毫不费力地得到放弃呼叫和其他故障等方面的客户层次数据。分析员也能容易地看到客户的购买历史，即使一些购买发生在商店里，一些通过邮寄订购目录，而有一些通过Web。对客户服务中心的每一次呼叫，同样可以容易地确定呼叫的持续时间及呼叫是否通过人工台转接或通过IVR(交互式语音应答)，在后一种情况下，通过语音提示，客户走过什么路径。最佳情况是，当需要的数据不太可用时，就会有一组人为它工作而使它可用，这可能意味着需要重新设计一份申请表格，重新改编自动交换机——或者简单地说，在最初就加载正确的数据。

16.2.2 将数据转化为可操作信息的技巧

理想的数据挖掘环境是由那些在数据处理方面技能高超的人组成的，数据挖掘只是通过他们对于商业运作方式及今后目标的深刻理解完成的。一个数据挖掘小组包括数据库专家、程序设计者、统计学专家、数据挖掘者和商业分析师，所有人一起工作确保商业决策以准确的信息为基础。该团队的人们有沟通技巧，能够把他们获知的所有内容传达到组织中的恰当部门，无论是市场营销、运作、管理或决策部门。

16.2.3 所有必需的工具

理想的数据挖掘环境包括足够的计算能力及数据库资源，以支持最详细层次的客户交易数据的分析，它包括能处理所有数据和由此创立模型集（model set）的软件。当然，它还包含一系列丰富的数据挖掘软件，以便应用第 5～13 章讲述的所有技术。

16.3 返回现实世界

我们从未见过上面描述的理想数据挖掘环境，读者不必吃惊。然而，我们已经合作的许多公司正在往这个正确的方向上努力，这些公司正在采取措施设法把自己转变成以客户为中心的组织。他们正在建立数据挖掘组，收集来自操作系统的客户数据，创建单一客户视图，其中许多已经初见成效，在收获实质上的利益。

16.3.1 建立以客户为中心的组织

本章首的乌托邦式幻想的第一个组成部分，是一个真正以客户为中心的组织。说到数据，建立以客户为中心的组织的最困难工作之一就是建立单个客户视图，在整个企业共享它，指导每次客户交互。与该挑战相对应的另一方面是，通过与客户交流的所有渠道，建立公司及其品牌的一个形象，这些渠道包括零售商店、独立经销商、Web 站点、客服中心、定向市场营销等。其目标不仅是做出更多明智的决策，而且以可测量的方式改进客户体验。换句话说，客户策略有分析和操作两种成分。本书更关心分析成分，但两者对成功都很重要。

提示： *建立以客户为中心的组织需要同时使用包含分析和操作成分的策略。尽管本书讨论的是有关分析的成分，但操作成分同样重要。*

建立以客户为中心的组织需要把来自多种来源的客户信息集中到单个数据仓库中，同时还要有一系列共同的定义和容易理解的商业过程一起来描述数据的来源。这种组合使我们可能定义一系列可用于所有组的客户度量及商业规则，来监控业务发展及测量市场条件变化和新措施的影响。

当然，客户信息的集中储存就是上一章中描述的数据仓库。如图 16-1 所示，在操作系统和数据仓库之间有双向交流。操作系统提供进入数据仓库的原始数据，而仓库反过来为操作系统提供客户评分、决策规则、客户片段定义及行动触发器。例如，一个零售 Web 站点的操作系统捕获了所有的客户订单，然后这些订单在一个数据仓库中被汇总。使用来自数据仓库的数据，建立关联规则（association rule），用于产生交叉销售建议，这些建议再返回到操作系统中。最终结果是：客户来到网站除订购了一条裙子外，还另外订购了几件紧身衣。

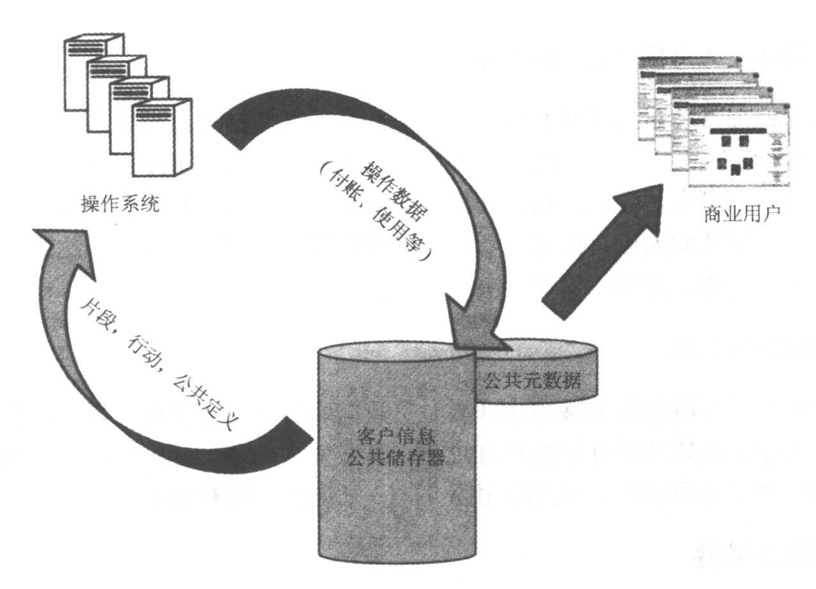


图 16-1 一个以客户为中心的组织需要集中的客户数据

16.3.2 创建单个客户视图

公司的每个部门应该能够访问客户的单个共享视图，呈现给客户的也应该是公司的单一形象。从实践的角度来说，这意味着共享单个客户的收益模型、一个单独的默认支付风险模型、单个客户的忠诚度模型，以及一些术语的共享定义，诸如客户启动、新客户、忠诚客户和有价值的客户。

对于这些术语，不同组有不同的定义是很自然的。在一个出版公司中，关于谁是最有价值的客户问题上，发行部门和广告营销部门有不同的观点，因为支付最高订阅价格的人不一定是对登广告者最感兴趣的人。解决方法是给每个客户分配一个广告价值和一个订购价值，使用像第4章中介绍的广告适应度的思想。

在另外一家公司，财务的风险管理组认为客户在最初4个月的保有期内是“新”的，在这个初始试用期内，任何推迟的付款被密切追踪。然而，客户忠诚组认为客户在最初3个月是“新”的。在这个欢迎期内，客户受到额外的照顾。那么，它到底属于哪一种：初期和谐的新关系还是试用阶段？如果在公司内部没有统一，顾客接收到的将是混合信息。

对于有几种不同业务的公司，问题会更微妙。同一个公司可能提供 Internet 服务及电话服务。当然，对于这两项服务会保持不同的支付方式、客户服务和操作系统。此外，如果ISP（因特网服务提供商）是新近被电话公司并购的，公司对现有电话客户和最近获得的因特网客户之间的重叠情况可能一无所知。

16.3.3 定义以客户为中心的度量标准

1929年9月24日，美国空军中尉 James H. Doolittle 创造了历史，他通过“盲飞”证明，利用最新发明的仪器如陀螺地平仪、方向回转仪、气压高度计等，即使把驾驶座舱用帆布引擎罩盖起来，也可以飞行一段精确的路程。在陀螺地平仪发明之前，飞行员飞入云层或

雾峰中时，常会迷失飞行方向。现在，得益于驾驶员座舱中那些仪表的帮助，我们可以在James中尉都会出现问题的恶劣天气中，从容地咀嚼椒盐卷饼，喝着咖啡茶，修改电子数据表（spreadsheet）。好的商业度量标准就像是保持一个大企业航向正确那么重要。

商业度量标准是告诉管理者在哪个方向上移动哪个控制杆的信号。选择正确的度量标准是重要的，因为该企业趋向于变为测量标准指出的方向。如果一个企业按照所拥有的客户数量考核自身，则该商业将趋向于签约新客户而不考虑保有期或潜在客户的未来收益前景。按照市场占有率考核自身的企业，将趋向于以其他的目标（如收益）为代价增加市场占有率。希望成为以客户为中心的公司所面临的挑战是提出现实的以客户为中心的度量标准。说公司的目标是增加客户忠诚度，这听起来很棒，但很难提出一个好方法对客户质量进行测量。仅仅持续一段较长时间是忠诚度的标志吗？或者忠诚度应当定义为能够抵抗来自竞争者的优惠服务？如果是后者，如何测量它呢？

即使表面上简单的度量标准，如流失或收益，也可能难以准确说明什么时候会发生流失：

- 电话服务实际不活跃的那一天？
- 客户最初表达不活跃意图的那一天？
- 不活跃后的第一个付账周期的后期？
- 当电话号码发放给新客户的那个日子？

上述的每一个定义在电话业务的不同部分中都扮演一个角色，对于有合同的手机用户，可能不存在这些事情。哪一种流失事件应该被看做是自发的？让我们来看一位客户的例子，他为了抗议恶劣服务而拒绝付款，最后被迫切断信号，这种流失是自发流失还是强制流失？那些自发停止又没有支付最终欠款的用户又如何呢？这些问题没有一个合适的答案，它们确实暗示着定义客户关系的微妙之处。

对于收益，哪些客户被认为是有收益的在很大程度上取决于成本是如何被分派的。

16.3.4 收集正确的数据

一旦正确定义了像忠诚度、收益及流失等，下一个阶段是确定需要的数据以便正确地计算它们。这不同于应用任何碰巧可用的数据简单地给出近似定义。记住，在理想的数据挖掘环境中，数据挖掘组有能力决定什么数据可用！

管理商业所需要的信息应该能够驱动添加新表和字段到数据仓库中。例如，一个以客户为中心的公司应该能够说出哪一个客户是有利可图的。在许多公司中这是不可能的，因为没有足够可用的信息能够敏感地分配在客户层次上的花费。我们的客户之一，一家无线电话公司，解决这个问题的方法是，通过编辑一个要求回答的问题列表，决定针对一个特定客户提供多少服务经费。然后他们确定回答这些问题需要哪些数据，并建立一个方案来收集这些数据。

这些问题很多，其中包含了下列各项：

- 客户每年呼叫客服中心多少次？
- 客户通过在线、支票还是信用卡支付账单？
- 客户花费在漫游上的时间比例是多少？
- 客户在哪些外部网络上漫游？
- 这些网络的签购成本是多少？
- 客户对客户服务中心的呼叫是由IVR还是由人工接线员处理？

回答这些成本相关的问题需要来自呼叫中心系统、账号系统及财务系统的数据。围绕其他重要度量标准的内容揭示了对呼叫明细数据、人口统计学数据、信用数据及Web使用数据的需要。

16.3.5 从客户交互到学习机会

以客户为中心的组织与它的客户保持一种学习关系，每一个与客户的交互都是学习的机会，是数据挖掘者和公司内部各种面向客户的组之间进行良好交流时能够被抓住的机会。

公司发生的几乎任何行为都会影响客户——价格的变化、新产品的引入、市场营销活动，都能被设计，所以它也是一个向客户进行更多学习的实践机会。这些实践的结果能够找到使数据进入数据仓库的方法，在那里可以进行分析。通常，行动本身是由数据挖掘提出来的。

例如，一家无线电话公司的数据挖掘结果显示，由于延迟支付而出现的暂停服务是*自发流失*（voluntary churn）和*强制流失*（involuntary churn）两者共同的预测器。延迟支付是一个不支付的预测器并不让人吃惊，但延迟支付（或公司对于延迟支付者的处理）是一个自发流失的预测器，则似乎需要更深入的调查。

这个观察导致这样的假说，暂停服务降低了客户对公司的忠诚度，而且，当有机会出现时，他们很可能会把业务迁到其他地方。信用卡署的数据清楚地表明，一些延迟支付者经济上有能力支付他们的电话账单。这提示我们进行一项实验：这些低风险客户应该与高风险客户区别对待，在终止他们之前，应该对他们的不良行为更耐心些，使用比较温和的方法劝说他们付账。一个对照实验测试了这一方法是否会提高客户忠诚度而不必让呆账提升。两个相似的低风险、高价值客户群得到了不同的对待，一个被作为"商业常规"处理，而另一个得到了比较亲切又比较温和的处理。在试验期结束的时候，基于保持及呆账比较这两个组，以便决定转换到新处理方式的经济影响。可以非常肯定地说，亲切、温和的处理方式对转变较低风险的客户被证明是值得的——增加付款比率，还稍微增加了客户的长期保有期。

16.3.6 挖掘客户数据

当每个客户的交互产生数据的时候，数据挖掘就有了无数的机会。可以挖掘购买模式和使用模式生成客户片段；挖掘响应数据以改进未来活动的目标；多重响应模型能够被结合生成最佳未来促销模型；生存分析可用于预测未来客户的流失；流失模型能够发现客户流失的风险；客户价值模型能够识别值得保持的客户。

当然，所有这些要求增设一个数据挖掘组以及支持它的基础结构。

16.4 数据挖掘组

数据挖掘组专门负责建立模型，使用数据了解关于客户的知识——与引导市场营销工作、设计新产品等相反。也就是说，这个组有技术职责而不是商业职责。

在公司层次上，我们已经看到数据挖掘组可以有几种不同的组织结构：

- 在公司外作为一个外包行为；
- 作为IT的一部分；
- 作为市场营销、客户关系管理或财务组织部分；
- 作为一个跨学科的小组而成员仍然属于各自的部门。

上述每一个结构各有其优缺点，下面将分别进行讨论。

16.4.1 外包数据挖掘

公司有多种理由考虑使用外包数据挖掘。对有些情况，数据挖掘只是偶尔需要，因此不值得投资建一个内部小组；另一些情况是，数据挖掘是一个当前正在成长的需要，但所需要的技术似乎与公司现有技术不同，在公司中白手起家创建这种专用技术会面临巨大的挑战。还有一些公司，它们的客户数据主体寄存在一个外部的卖主那里，感觉分析过程应该在数据所在地进行。

1. 外包偶然建模

一些公司认为，他们对建立模型及使用数据了解客户没什么需求。这些公司通常分为以下两种类型：第一种是公司客户很少，要么是因为公司太小，要么是因为每个客户太大。例如，一个典型银行的私人银行业务组可能只会为数千客户服务，而账号代理人了解他们的客户。在这样的环境下，数据挖掘可能是多余的，因为客户关系是如此密切。

不过，即使在这个环境中数据挖掘也能扮演一个角色。特别地，数据挖掘能够使我们了解最佳实践并传播它们。例如，私人银行中的一些职员可能以某些方式（如保留客户，鼓励客户推荐朋友、家庭成员、同事等）把工作做得更好，这些职员可能有最佳实践方法需要在整个组织中传播。

提示： 如果公司拥有与客户维持深入及长期关系的尽职职员，则数据挖掘可能是不必要的。

对于在新兴市场中的快速成长的公司，数据挖掘似乎也不那么重要。在这种情况下，客户获取（customer acquisition）驱动商业的发展，广告（而不是定向市场营销）是吸引新客户的主要方法。数据挖掘在广告上的应用是有限的，而且在这一时期的发展中，公司尚未把重心集中在客户关系管理和客户保持上。对于他们所做的有限的定向市场营销，外包建模通常就已经足够了。

无线通信、有线电视及互联网服务提供商全部经过了指数生长期，只不过这种增长最近已经结束，市场已经成熟（而在此之前，有线电话、人寿保险、目录销售及信用卡经历了相似的周期）。在初始成长期，了解客户可能不值得投资——加设一个发射塔、交换机或其他类似的东西都可以提供较好的回报。最后，业务与客户基础增长到一个点，此时了解客户具有越来越大的重要性。根据我们的经验，公司最好尽早朝着了解客户的方向开始，而不是等到需求达到临界点时才开始。

2. 外包正在进行的数据挖掘

即使当一家公司已经认识到数据挖掘的必要性，仍然有可能外包数据挖掘。当公司建立在以客户获取为基础的时候尤其如此。在美国，信用卡公司及家庭数据供应者很乐意提供建模，利用它们出售的数据进行增值服务。也有直接的营销公司负责从邮寄列表到结果的每件事情——包括对客户的实际产品递送。这些公司时常提供外包数据挖掘。

外包对于公司经济是有利的，问题是对客户的深入了解也是外包性的。一家仰赖外包客户分析的公司存在这样的风险：在公司和厂商之间对客户的理解可能会有遗失。

例如，一家公司利用直接邮寄作为获取客户的重要手段，而且把直接邮寄响应建模（response modeling）工作外包给邮寄列表厂商进行。在大约 2 年的过程中，公司曾经有几个

直接邮寄管理人员，而且对这一渠道的关注在减少。没有人认识到，直接邮寄一直在增加获取量，但这种获取却算到了其他渠道的头上。直接邮寄的表格可能已被填写，然后通过邮件寄回，这种情况下新的获取是计算到直接邮寄账上的。但这些邮件也包含了公司的网址和一个免费的电话号码。许多收到直接邮寄的潜在顾客发现通过电话或在 Web 上回应更方便，而通常忘记提供用于将其识别为直接邮寄潜在顾客的专门代码，随着时间的过去，归因于直接邮寄的响应减少，从而用于直接邮寄的预算也减少。直到最后，当减少直接邮寄导致其他流通渠道的响应减少的时候，公司才认识到忽视这一响应结果已经导致他们做出了一个低于最佳效果的商业决策。

16.4.2 内部数据挖掘

建模过程产生的不止是模型和得分，也产生洞察力，这些深入了解通常来自数据探测和数据准备阶段，这是数据挖掘过程的一个重要部分。正由于这个原因，我们认为任何正在出现数据挖掘需求的公司应该发展内部的数据挖掘组，以便在公司内部探究数据。

1. 建立一个多学科交叉的数据挖掘组

一旦做出决策，在公司内部产生对客户的理解，问题就出现了。在一些公司中，数据挖掘组没有永久的场所。组成员聚在一起，在本职工作之外完成数据挖掘，本质上看，这样的安排似乎是暂时的，且通常它是一些紧急需求（如需要了解突然发生的客户违约高潮）的结果。当这样的一个组存在的时候，可能是非常有效的，但是不可能持续很长时间，因为一旦有新的任务，这些成员就会被召回到他们正常的工作中。

2. 在 IT 组织中建立一个数据挖掘组

一个可能的场所是在系统组中，因为这个组常常负责存储客户数据且运行面向客户的操作系统。因为数据挖掘组是关于技术的，并且需要数据的存取和强有力的软件及服务器，所以 IT 组织似乎是一个很自然的选择。事实上，分析可以视为提供数据库和存取工具及维护这种系统的进一步延伸。

作为 IT 组织的一部分，还有这样的优势：在需要的时候，数据挖掘组能够接触到硬件和数据，因为 IT 组织具有这些技术上的资源和对数据的访问权。此外，IT 组织是一个在许多企业单位拥有客户的服务组织。事实上，作为数据挖掘“客户”的企业单位或许已经习惯于仰赖 IT 组织给出的数据和报告。

另一方面，IT 组织有时与推动客户分析的商业问题稍稍有点距离，因为对商业问题轻微的误解能够导致无用的结果，所以让来自企业单位的人紧密地参与到以 IT 人为基本成员的数据挖掘计划是非常重要的。

3. 在企业单位中建立一个数据挖掘组

把数据挖掘组与存放数据和计算机的地方结合在一起的另一种方法是，把它与要解决的问题放在一起。这通常指的是市场营销组、客户关系管理组或财务组。有时会有几个小的数据挖掘组，每个企业单位都有一个，一个在财务组建立信用风险模型和采集模型，一个在市场营销组建立响应模型，一个在 CRM 组创建交叉销售模型及自发流失模型。

这一方法的优点和缺点正好与置入 IT 组织中的数据挖掘相反。企业单位对于自身的商业问题都很了解，但可能仍不得不依赖 IT 组织作为数据及计算处理源。虽然每一种方法都可能成功，但是总的来说，我们还是希望数据挖掘被置于企业的中心。

16.4.3 数据挖掘组成员需要具备的条件

最好的数据挖掘组通常选择复合型人才，因为数据挖掘作为一个单独的活动，存在的时间并不长，只有少数人可以声明接受过训练，成为数据挖掘者。有的数据挖掘者过去一直是物理学家，有的过去是地质学家，有的是计算机科学家，有的过去是销售经理，有的是语言学家，还有的是统计学家。

这使得数据挖掘组的午餐时间交谈非常有趣，但是它不会给雇佣经理提供更多的指导。使好的数据挖掘者胜于普通人的要素是很难教会的，也不会自动生成，那就是：良好的直觉，如何从数据中巧妙地获取信息的感觉，以及自然的好奇心。

没有任何一个人可以具备完成一项数据挖掘计划需要的所有技能。在他们之中，组成员应该覆盖下列技能方面：

- 数据库技能（SQL，如果数据存储在关系数据库中）；
- 数据转换和编程技能（SAS，SPSS，S-Plus，PERL，其他编程语言，ETL工具）；
- 统计学；
- 机器学习技能；
- 相关行业知识；
- 数据可视化技能；
- 访问及需求收集技能；
- 展示、写作和沟通技能。

一个新成立的数据挖掘组应该包含以前已经做过商业数据挖掘的人——最好是在相同的行业中。如果需要，这个专家可以由外部顾问公司提供。

16.5 数据挖掘基础设施

在认为数据挖掘只是一个探索性活动的公司中，数据挖掘可以在几乎没有基础设施的情况下完成。一个台式机工作站加上一些数据挖掘软件以及对企业数据库的访问可能就是足够的，然而，当数据挖掘成为企业的核心内容时，数据挖掘的基础设施一定要非常地强健。在这些公司中，要随时利用新的模型得分更新客户简档，这种更新或者是定期的（如按照进度表每月一次），或者是在某些情况下，对每一次的新交易进行更新。它已经成为数据仓库常规生产过程的一部分。数据挖掘的基础设施必须在发展模型的探索领域和模型被评分、市场营销活动在进行的生产领域之间提供一架桥梁。

可用的数据挖掘环境一定能够支持下列各项任务：

- 从许多来源访问数据以及把数据汇集在一起成为一个数据挖掘模型集中的客户特征标识（customer signature）的能力；
- 根据需要从模型库中使用已经创建的模型给客户评分的能力；
- 按时间处理数以百计模型得分的能力；
- 按时间处理得分或开发出的数以百计模型的能力；
- 在客户保有期中的任何一点上重新建立客户特征标识的能力，比如一次购买或其他有趣事件恰巧发生之前；
- 追踪模型得分随时间变化的能力；

- 给数据仓库及需要它们的其他应用软件发布得分、规则及其他数据挖掘结果的能力。

数据挖掘的基础设施从逻辑上（通常实体上也是）可以分为两个部分，支持两个完全不同的活动：挖掘及评分。每个任务表现出一系列不同的需求。

16.5.1 挖掘平台

挖掘平台支持数据处理软件，也支持本书中描述的具体表达数据挖掘技术的数据挖掘软件、可视化及显示软件，以及能够使模型公布环境得分的软件。

虽然我们已经提到了一些综合性问题，下面这些方面还应考虑：

- 在客户/服务器的分层中，软件应当安装在哪里？
- 数据挖掘软件需要自己的硬件平台吗？如果需要，它会在各种混和的系统中引入一个新的操作系统吗？
- 为了与软件包沟通，什么样的软件将被安装到使用者的台式机上？
- 需要什么样的附加网络、SQL 网关及中介软件？
- 数据挖掘软件为报告和图形软件包提供良好的接口吗？

挖掘平台的目的是支持数据探测、挖掘及建模，系统构思时应该把这些活动记在头脑中，包括此项工作需要更多的处理过程及计算能力这一事实。数据挖掘软件厂商应该能够提供关于适合预期的数据库大小及使用模式的数据挖掘平台的详细说明。

16.5.2 评分平台

在挖掘平台中发展起来的评分平台模型被应用于客户记录，以便创建用于确定未来措施的得分。通常，评分平台可能就是客户数据库本身，它可能是一个关系数据库，运行于并行硬件平台上。

为了给一个记录评分，该记录必须包含（或者评分平台必须能够计算出）送入模型的相同特征。这些模型使用的特征很少是未加工的数据原始形式。通常，新的特征是以不同的方式结合现有的变量创建的，例如取其中一个对另外一个的比，并进行分箱、求和及求平均值等。无论进行了哪些计算，创建模型时所使用的特征现在一定完成了对每个记录给出得分的工作。因为可能有数以亿计的交易记录，如何完成这件事情是很重要的。当数据量很大的时候，数据处理面临的挑战也很大。

直到得分被放入一个易于被软件存取的客户数据库，评分才会完成。这个得分常常用于在营销活动中选择包括哪些客户。如果作为模型输入的 Web 日志、呼叫明细或销售点扫描器数据处于一个系统的固定文件中，而客户营销数据库归在另外一个系统上，但按照不同的日期，这两个库都是准确的，这可能也是数据处理的一个挑战。

16.5.3 一个产品数据挖掘结构实例

对于常规的把数据挖掘和评分整合到操作环境方面，Web 零售比大多数产业走得更远。许多 Web 零售商利用每一次交易更新客户简档，应用模型得分确定该展示什么及推荐什么。这里描述的结构来自 Blue Martini 公司，该公司提供为挖掘作准备的零售网站软件。它提供了数据挖掘如何成为公司运转的组成部分的例子，这个例子并不局限于 Web 零售商，许多公司可以受益于一个相似的结构。

1. 结构概览

Blue Martini 结构被设计成支持市场营销者、商人以及众多数据挖掘者的不同需求。如图 16-2 所示，对于三种不同类型的用户，它有三个模块。对于商人，这个结构支持多重产品分层和工具，以控制收集和促销。对于市场营销人员，有进行对照实验（controlled experiment）以追踪各种信息及市场规则有效性的工具。对于数据挖掘者，有完整的建模软件，可以使他们从几十个不同的服务器和应用程序记录通过手工处理来创建客户特征标识的工作中解放出来。这就是 Ralph Kimball 和 Richard Merz 所称的网上数据仓库，它是由几个特殊目的的数据集市以不同的方式建立的。所有仓库都使用了共同的字段定义、共享元数据库。

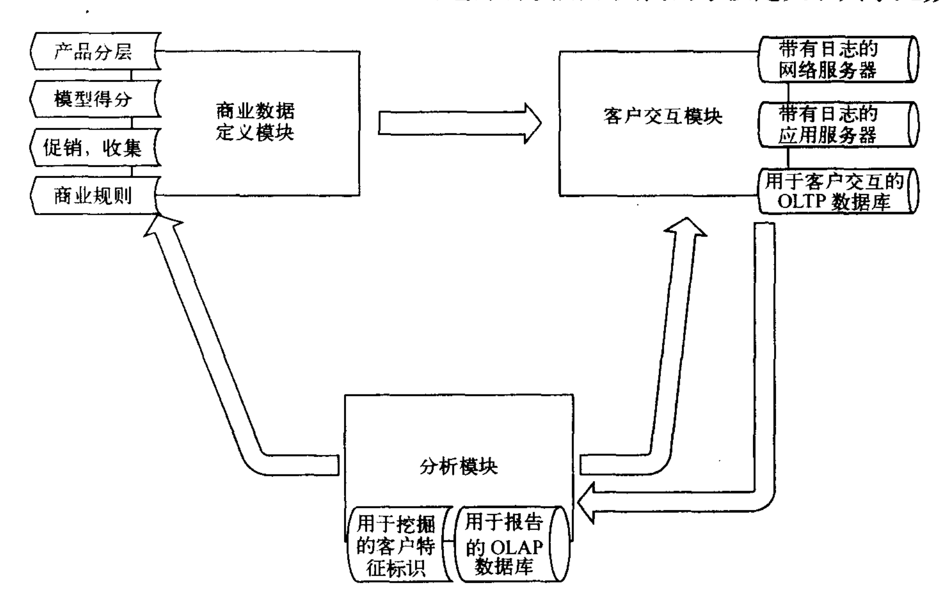


图 16-2 Blue Martini 为由数据挖掘驱动的 Web 零售商提供了一个 IT 结构的好例子

Web 商店的客户与按照需要从包含产品信息和页面模板的数据库生成的页面进行交互。页面中的内容是按规则生成的，这些规则中有些是由管理员输入的商业规则，另一些则是自动生成的，然后经过专业销售人员的编辑。

由一个数据库生成页面有许多优点。首先它可以加强整个 Web 站点视觉和感觉的一致性，这样的标准界面帮助客户在该站点内畅行无阻。使用数据库可以提高全局性变化的速度，比如为更新价格进行降价促销。另外的一个特点是以不同的语言和币种存储模板的能力。因此，站点能够为不同国家的用户定制。从数据挖掘观点看，最主要的优点是所有的客户交互都可记录到数据库中。

用户交互通过一系列数据集市进行管理。报告和挖掘处于客户行为数据集市的中心，它包括来自用户交互、产品和商业规则等数据集市的信息。从交易数据创建客户特征标识所需要的复杂提取与逻辑转换工作是系统的一部分——这简化了任何一个曾经尝试处理 Web 记录以获取客户信息的人的工作。

2. 客户交互模块

这一结构包含所需要的数据库和软件，以支持交易、客户交互、报告和挖掘，也支持个性化的以客户为中心的市场营销。Blue Martini 公司的系统有三个主要的模块，每个模块有

自己的数据集市。这些资源库保存着下列各项记录：

- 商业规则
- 客户及来宾交易
- 客户行为

客户行为数据集市（如图16-2所示，它是分析模块的一部分）由来自客户交互模块的数据作为输入，它反过来又为商业数据定义模块及客户交互模块提供规则。

商品交易信息如产品层级、分类（为交易目的而被聚集在一起的产品族）和价格列表等保存在商业规则数据集市中，同样保存在其中的还有如Web网页模板、图像、声音和视频片段。商业规则包括问候指定客户的个性化规则、促销规则、交叉销售规则等。为一个零售站点进行数据挖掘的工作很大一部分是用于产生这些规则。

客户交易模块是系统通过处理所有客户交易而直接接触客户的部分。客户交易模块负责维持用户的生存期及环境。这一模块实现了实际的Web商店，并且收集以后分析可能用到的任何数据。客户交易数据集市记录的商业事件示例如下：

- 客户把一个项目加入到购物篮中；
- 客户开始付账过程；
- 客户完成付账过程；
- 交叉销售规则被触发，给出推荐商品；
- 紧接着是被推荐物品的链接。

客户交互模块可以通过加入对照群组和保持多重规则的方法来支持市场营销实验。它有其服务对象的详细知识，能追踪没有被Web服务器记录的许多事情。客户交互模块收集的数据使我们可以观察产品和客户随时间变化的情况。

3. 分析模块

支持客户交互模块的数据库，像大多数的在线交易处理系统一样，是一个设计用于支持快速交易过程的关系数据库。专门用于分析模块的数据必须被提取并转换为支持挖掘及报告的适当结构。数据挖掘要求有一个固定的特征表格，其中对每个待研究的客户或项目有一行信息。这意味着要求有某些变体，它们收缩了产品分层，以至于会出现这样的情况，比如，同一项交易可能产生一个标记表明该客户购买了法国酒，产生另一个标记说他（或她）购买了一种产自Burgundy的酒，而第三个标记则可能指出该酒来自Burgundy的Beaujolais区。其他数据必须从包含每位客户多重交易记录的订货档案、账单档案及购物时间段中积累得到。由此得到的典型数值包括按种类计的全部花费、平均订购数、该客户的平均订购和均值平均订购之间的差，以及客户自上次购买至今的天数。

报告由一个多维的数据库实现，该数据库允许在不同层次回溯查询。数据挖掘及OLAP都是分析模块的一部分，尽管它们回答的是不同类型的问题。OLAP查询通常用来回答如下问题：

- 销量最好的产品是什么？
- 销量最差的产品是什么？
- 浏览最多的是哪个页面？
- 按照商标名计，变换速度怎样？
- 按照访问计数，指向最多的网址是哪一些？

- 按照销售总额计，指向最多的网址是哪一些？
- 有多少客户抛弃了购物篮？

数据挖掘用于回答更复杂的问题，比如：

- 大宗买家的特征是什么？该用户适合这个简档吗？
- 应该给这位客户提供什么样的促销手段？
- 这个客户在1个月内再回来购买的可能性有多大？
- 什么样的客户是我们应该担心的，因为他们最近没有访问本站点？
- 什么样的产品与花钱最多的客户相联系？
- 什么样的产品正带动着其他什么产品的销售？

在图16-2中，标有“建立数据仓库”的箭头连接客户交互模块与分析模块，代表在数据挖掘或报告真正完成之前必须进行的所有转换。另外多出的两个箭头标记为“展开结果”，显示了分析模块的输出结果，反馈回商业数据定义模块和客户交互模块。而另外一个箭头标记为“阶段数据”，显示了商业规则如何插入到客户交互模型的商业定义模块中。

这一系统结构最吸引人的是它驱动数据挖掘良性循环的方式。它允许将数据挖掘发现的新知识直接反馈到与客户交互的系统中。

16.6 数据挖掘软件

从本书的第1版问世以来，数据挖掘领域方法的最大改变之一就是，数据挖掘软件产品变成熟了。强大的功能、可用性及可扩展性都有了很大的提高。可能有所下降的就是数据挖掘软件商数量，因为小作坊式的软件公司已经被更大、更让人接受的公司挤跨了。正如本书第1版所说，在一本书中比较特定产品的优点，以便使当前产品在货架期之外仍然保持有用是不合理的。虽然产品正随时间而改变——乐观地说是改善，评价的标准一直没有改变：价格、有效性、可扩展性、支持情况、厂商关系、兼容性以及可以方便地把所有因素集成在一起选择的过程。

16.6.1 所应用的技术范围

目前必须清楚的是，没有单一的数据挖掘技术可适用于所有的情形。神经网络、决策树、购物篮分析、统计学、生存分析、基因算法、基于存储的推理（memory-based reasoning)、链接分析和自动聚类探测（automatic cluster detection）都有一席之地。正如案例研究所展示的，将这些技术中的两种或多种结合在一起所达到的效果远远超过任何单一方法，这并不稀奇。

首先要保证所选用的软件的能力足以支持企业的数据和需求目标。让软件稍微超前于分析者的能力是一个好主意，这样人们就可以试验他们可能想不到去尝试的新事物。在一个工具包中有多种可用的技术是有用的，因为它使结合和比较不同的技术变得比较容易。同时，有几种不同的产品对于一个较大的企业来说也是合理的，因为不同的产品有不同的力量——即使它们支持同样的基本功能。某些对于呈现结果是比较好的，某些对于给出得分是比较好的，另一些对新手用户更直观。

评估将要进行的数据挖掘任务的范围，决定使用哪一种数据挖掘技术将会是最有价值的。如果你心里已经想到单个应用软件，或几个密切相关的应用软件，那么你可能选择这种单一技术并坚持使用它。如果你正在建立数据挖掘实验室环境，以处理范围广泛的数据挖掘

应用，那么你可能想要寻找一个配合良好的工具套装。

选择数据挖掘软件需要考虑的问题

下列问题是为你的公司设计的，可以帮助选择正确的数据挖掘软件。我们给出的问题是一个无序的列表，你应该做的第一件事是依照自己的优先次序进行排序。对于不同的案例，这些优先次序肯定会不同，这就是我们为什么没有尝试事先将它们排序的原因。例如，在某些环境中，有一个确定的标准硬件供应者，独立平台不是一个问题；而在其他环境中，人们最为关心的是，如此不同的部门能够使用这个软件包，或者期望在硬件方面将来会有一个变化。

- 由厂商提供的数据挖掘技术的应用范围是什么？
- 对于数据的大小、用户的数目、数据中的字段数目以及它使用的硬件来说，产品的可扩展性如何？
- 该产品对数据库和文档提供透明的访问方式吗？
- 产品能提供多种层次的用户界面吗？
- 产品对于它产生的模型能提供可理解的解释吗？
- 该产品支持图形、可视化及报告工具吗？
- 产品与环境中的其他软件（如报告软件包、数据库等）交互情况良好吗？
- 产品能够处理不同的数据类型吗？
- 产品是否便于存档？使用简单吗？
- 支持、训练及咨询方便吗？
- 产品适应现有计算环境的程度如何？
- 厂商有可信的介绍人吗？

一旦你确定了上述哪一个问题对你的组织最重要，通过与软件厂商面谈，或者从一个独立的数据挖掘顾问处获得帮助，就可以利用你选出的问题来评估候选软件包。

16.6.2 可扩展性

当被处理的数据量大而复杂的时候，数据挖掘会提供最佳帮助。但是，数据挖掘软件可能在小的样本数据集中演示的，所以要确定所考虑的数据挖掘软件能够处理预期的数据量——然后可能更多地要考虑将来的数据成长（数据不会随着时间变得更小）。数据挖掘的可扩展性对以下三个方面是很重要的：

- 将数据转换成客户特征标识需要许多输入/输出和计算能力；
- 创建模型是一项重复性的和投入非常大的计算；
- 评分模型需要复杂的数据转换。

为探究和转换数据，最方便可用的可扩展软件就是关系数据库。它们是专门设计的，可以充分利用多处理器和多磁盘等特点来处理单一数据库查询。另一类软件，用于创建数据库的提取、变换和装载（ETL）工具对于数据挖掘也可能是可扩展和有用的。然而，大多数的程序语言不能扩展，它们只支持单一处理器和单一磁盘来处理单一任务。当有许多数据需要结合的时候，处理这些数据最容易的扩展办法时常在这个层次被发现。

建立模型和探究数据需要运行足够快且能够在足够大量的数据上运行的软件。一些数据挖掘工具只能作用于内存中的数据，因此，数据的容量被有效内存所限制，它所具有的优点是算法运行得更快，但存在局限。实际上，当可用内存以兆字节计算的时候，这曾经是一个

问题，但在典型的可用内存为千兆的工作站上改善了这个问题。通常，数据挖掘环境把多用户数据挖掘服务器放在靠近数据的一个强有力的服务器上，这是一个好的解决办法；当工作站变得更强的时候，在本地建立模型也是可行的解决方法。在这两种情况中，目标都是在合理的时间内运行模型中的数十万行或数百万行。数据挖掘环境应该鼓励使用者了解和探究数据，而不是花费精力降低样本大小以使它适合。

因为评分环境需要转换数据同时运行模型，所以它通常是最复杂的——最好有最小量的用户介入。也许最好的解决办法是在数据挖掘软件既能读又能写到关系数据库的时候，这样就可能把数据库用于可扩展的数据处理，把数据挖掘工具有效地用于建立模型。

16.6.3 评分支持

当数据挖掘用于开发评分模型的时候，对数据库的写入和读取能力是非常重要的。模型可能是使用从主数据库抽取的样本来建立的。但一旦被建立起来，模型将会用于给数据库中的每个记录评分。

响应模型的价值随时间而减少。理想的情况是，一项活动的结果应该及时被分析以便影响下一项活动。但在许多组织中，模型建立的时间与它被用于对数据库评分的时间之间有一个长长的滞后，有时这一时间长度甚至为数个星期或数个月。这种延迟是由于把评分模型转换成可用于数据库的形式这一困难所引起的，因为评分模型时常是在与数据库服务器不同的计算机上发展出来的。这种转换可能包括解释数据挖掘工具的输出结果，写出一个具体表示构成模型规则的电脑程序。

当数据库实际上被储存在第三方设备（如表处理机）的时候，问题将会更糟糕，因为表处理器不可能接受 C 源代码形式的神经网络模型作为对一个列表选择请求的输入。建立一个统一的模型开发和评分构架需要付出大量精力，但是如果为大型数据库评分对于商业是一项重要请求，这种努力是有回报的。

16.6.4 用户界面的多种层次

在许多组织中，有几种不同的用户团体使用数据挖掘软件。为了适应他们各有差异的需求，工具应该提供一些不同的用户界面：

- 为偶尔使用的用户准备一个图形用户界面（graphical user interface，GUI），这种界面对数据挖掘参数设有合理的默认值；
- 针对更熟练用户的高级选项；
- 以批量模式（它可能是由一个指令行界面提供的）建立模型的能力；
- 应用程序编程接口（API），以便预言性建模工作能被内置到应用程序中。

数据挖掘工具的 GUI 不仅能够让使用者容易地建立模型，而且应该被设计成鼓励最佳实践，比如确保模型评估在一个保留集（hold-out set）上执行，确保预言性模型的目标变量来自比输入更晚的一个时间帧。用户界面应该包含一个帮助系统，给出相关的帮助。用户界面应该提供合理的默认值，比如，支持分裂一个决策树所需要的最少记录数，或者是为改善偶然用户成功机会的神经网络隐藏层结点数目。另一方面，界面应该可以让更熟练的用户改变默认值，高级用户应该能够控制潜在的数据挖掘算法的每一个方面。

16.6.5 可理解的输出

工具在解释自身的程度上差别非常大，规则生成器、树可视化、Web 图表及关联表格均能提供帮助。

一些厂商常把重点放在数据和规则的可视化表示方面，提供三维空间数据的地形图、地理信息系统（geographic information system，GIS）和聚类图表，以帮助理解复杂的关系。很多数据挖掘工作的最终目的是对管理进行报告，而图形对于非技术用户信服数据挖掘结果的力量不应该被低估。数据挖掘工具应该很容易地将结果输出到普遍使用的报告分析软件包（如 Excel 和 PowerPoint）中。

16.6.6 处理各种数据类型的能力

许多数据挖掘软件包对能够被分析的数据类型有所限制。在购买一个数据挖掘软件包之前，要查明它是否能够处理你想用的各种不同的数据类型。

一些工具对于用分类变量（如模型、类型、性别）作为输入变量有一定的困难，需要用户把它们转换成一系列的是/否变量，每个可能的类对应一个变量。其他的工具能处理取少数几个值的分类变量，但当面对太多值时会崩溃。对于目标字段来说，一些工具能处理二进制分类任务（好/坏），但对预测可以取几种数值的分类变量值有一定的困难。

市场上的一些数据挖掘软件包需要由用户把连续变量（收入、里程、余款）分解到相关范围。产生关联规则的工具尤其如此，因为这些工具需要一定数目的相同组合值，以便识别出一个规则。

大多数的数据挖掘工具不能够处理文本，虽然支持文本的工具已经出现。如果数据中的文本串是标准化代码（状态、零件号码），这没有任何问题，因为字符码能够容易地转换为数值型或分类型变量。但有时应用程序需要分析自由文本的能力，一些更高级的数据挖掘工具包已经开始提供对这些能力的支持。

16.6.7 文档及简单使用

一个设计良好的用户界面应该能够立刻开始挖掘过程，不过掌握工具需要花时间学习。和任何复杂软件一样，好的文档编制能够说明成功和挫败之间的差别。在对一个工具做出选择之前，请查看使用手册。产品说明文档应该完整描述使用的算法，不应该仅仅是对工具的操作进行说明。企业不应该基于没有弄明白的技术进行决策。依赖于任何所有者及未知“秘密技术”的数据挖掘工具不是好的选择。

16.6.8 对新手和高级用户的培训、咨询和支持

将不熟悉的数据挖掘技术引入一个企业并不是件容易的事，在把任务交给一个工具之前，需要从工具厂商或第三方确定是否存在有效的用户培训及使用咨询。

如果厂商较小且离你的数据挖掘工作位置距离遥远，客户支持可能是不方便的。因特网已经缩小了这个星球，所以与每个供应商的联系实际上仅仅是几个电脑按键，但是它没有改变人类晚上睡觉白天工作的习惯，时区仍然是有关系的。

16.6.9 卖方可信度

除非你已经熟悉厂商，否则应该想办法知道一些关于它的历史记录和未来前景。要试着与已经使用该厂商软件的用户交谈，以此作为参考，以证实厂商在产品说明书中声称的内容。

不是说仅仅因为厂商是新的、小的或者距离很远，就不应该从这样的公司购买软件，数据挖掘仍然处于商业决策支持技术的前沿，通常是那些小的、刚刚起步的公司最先了解新技术的重要性，并成功地把它们引入市场。较小的公司时常提供更好、更热情的支持，因为回答问题的人很可能也是设计和创建产品的人。

16.7 小结

理想的数据挖掘环境包含以客户为中心的企业文化和支持它的所有资源。这些资源包括数据、数据挖掘者、数据挖掘基础设施和数据挖掘软件。在这种理想的数据挖掘环境中，对良好信息的需求是企业文化中根深蒂固的东西，操作规程的设计始终把收集好的数据的需求放在首位，而且数据挖掘的需求引导企业数据仓库的设计。

建立理想的环境并不是件容易的事。要建立以客户为中心的组织，最艰难的工作是改变文化，如何完成这个过程已经超出本书的讨论范围。从纯数据的角度看，第一个阶段是创建一个单一客户视图，它包含了公司拥有的与该客户所有渠道的关系；下一个阶段是创建以客户为中心的度量，用于追踪、建模和报告。

只要有可能，客户间的交互就应该变成学习机会。尤其是，市场营销沟通应该作为对照实验。这些实验的结果可以当作数据挖掘模型的输入，用于寻找目标、交叉销售及客户保持。

有几种方法可以将数据挖掘融入到公司的销售和客户关系管理活动中。对于那些偶然有建模需求的公司，外包数据挖掘是可行的。当对数据挖掘有不断增长的需求时，最好是在公司内部完成。这样，在挖掘期间产生的深入了解就掌握于公司手中，而不是在外部卖主那里。

一个数据挖掘组可以在公司组织的几个位置中获得成功，若把这个组定位于 IT 组织，就把它放在靠近数据及技术资源的地方；若定位在一个企业单位内，就把它放在接近商业问题的地方。无论哪一种情况，在 IT 组织及企业单位之间都应有一个良好的沟通。

为数据挖掘环境选择软件很重要，然而，数据挖掘组的成功更多地依赖于优秀的程序和优秀的人员，而不是他们的台式机中的特殊软件。

第 17 章 为挖掘准备数据

半透明琥珀色的液体——汽油是支撑运输业的动力，它几乎不能与从油井抽出的黑色胶粘石油相提并论。这两种液体之间的差别是经过从原材料蒸馏有用产品的若干精炼步骤产生的结果。

数据准备是一个非常类似的过程。其间，原始数据来自于操作系统，其中的数据以古怪的商业规则与系统增强和修复的分层形式存在，时常堆积如山。数据中的字段用于多种目的，数值渐渐变得过时无效。人们以发展的眼光不断修复错误，因此，解释随时间变化。准备数据的过程就像炼油。有价值的东西潜藏在操作数据的淤泥中。一半的工夫是精炼；另一半是将它的能量转化成有用的形式，即靠汽油驱动引擎。

数据增值是现代商业的特征。挑战是使数据的存在有意义，精炼数据，以便数据挖掘引擎能抽取数值。挑战之一是数据的绝对量。客户可能一年几次致电呼叫中心，每月一次支付账单，每天一次开启电话，一天几次打出和接收电话。其间，数十万或数以百万计的客户在产生数以亿计的行为记录。既使在今天的计算机上，数据处理的量也是相当巨大的。幸运的是，计算机系统已经变得足够强大，真正的问题是，要有购买硬件和软件的适当预算。从技术来看，处理如此海量的数据是可能的。

数据的形式多种多样，来源于多个系统，存在类型各异。数据总是杂乱无章、不完全的，有时是无法理解和不兼容的。唉！这就是现实世界。并且对数据挖掘来说，数据仍然是原材料。汽油开始以粘稠物质的形态存在，与杂质混合在一起。只有经过不同阶段的精炼，原材料才被转变成有用之物——无论是清澈的汽油、塑料还是化肥。正如最有力的引擎不能够使用原油作为燃料一样，最有力的算法（数据挖掘引擎）不可能在尚未准备好的数据中发现重要的模式（pattern）。

经过一个多世纪的实践，炼油的步骤已经完全清楚——比数据准备的过程要清楚得多。本章通过举例说明一些基于经验的指导方针和原则，使准备过程变得更加有效。首先讨论准备好的数据看起来应该像什么样子，用它来描述客户特征标识（customer signature）。然后从数据类型和列角色的角度，详细研究数据实际上是什么样子。由于成功的数据挖掘的主要部分在于衍生变量（derived variable），与此相关的概念在本章都给出了详细介绍。本章的结束部分讨论脏数据和缺失值带来的困难，以及在大量商业数据上存在的计算挑战。

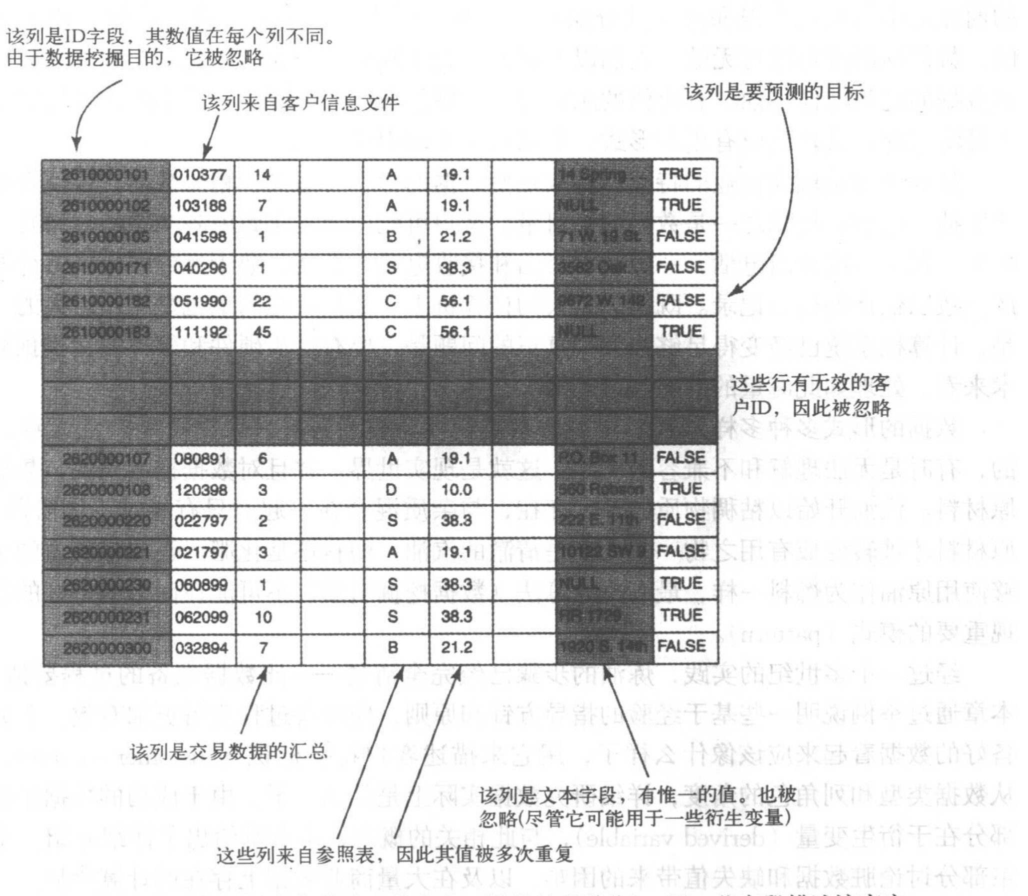
17.1 数据应该像什么

我们先讨论数据应该像什么。所有的数据挖掘算法要求输入是以表格的形式，即类似电子数据表和数据库中常见的行和列。然而，与电子数据表不同，这里的每个列对所有的行而言必须代表相同的意义。

一些算法要求数据具有特别的格式。举例来说，购物篮分析（已在第 9 章中讨论）通常只考察在任何给定的时间购买的产品。同样，链接分析（参见第 10 章）需要记录之间的参照以便连接它们。然而，大多数算法，特别是决策树、神经网络、聚类和统计回归都使用称为客户特征标识的特定格式的数据。

17.1.1 客户特征标识

客户特征标识是客户行为的快照，捕获客户当前属性和随时间的行为变化。和支票上的客户特征标识一样，理论上每位客户特征标识是惟一的——捕获个体的独特特征。然而不像支票上的客户特征标识，这里的客户特征标识是用于分析，而不是身份识别（identification）。事实上，与表示一个家庭、个人或账号的表面上看似随机的数字串相比，客户特征标识通常没有更多的识别（identifying）信息。图 17-1 显示，客户特征标识只是代表客户和任何对数据挖掘可能有用的简单数据行。



2610000101	010377	14		A	19.1		14 Spring ...	TRUE
2610000102	103188	7		A	19.1		NULL	TRUE
2610000105	041598	1		B	21.2		71 W. 19 St	FALSE
2610000171	040296	1		S	38.3		3562 Oak	FALSE
2610000182	051990	22		C	56.1		9672 W. 142	FALSE
2610000183	111192	45		C	56.1		NULL	TRUE
2620000107	080891	6		A	19.1		P.O. Box 11	FALSE
2620000108	120398	3		D	10.0		560 Robson	TRUE
2620000220	022797	2		S	38.3		222 E. 11th	FALSE
2620000221	021797	3		A	19.1		10122 SW 9	FALSE
2620000230	060899	1		S	38.3		NULL	TRUE
2620000231	062099	10		S	38.3		RR 1729	TRUE
2620000300	032894	7		B	21.2		1920 S. 14th	FALSE

图 17-1 客户特征标识的每行代表一位客户（数据挖掘单位），用一些字段描述该客户

也许不幸的是，没有大型数据库拥有现成的最新客户特征标识，可以直接用于所有的建模应用。这类系统初看可能非常有用。然而，这种系统缺乏的是机会，因为建模工作需要了解数据。虽然有些客户特征标识对一些应用工作良好，但是没有单一的客户特征标识能对所有建模工作都起作用。

在客户特征标识中的“客户”是数据挖掘的单位。本书主要关注客户，因此，典型的数据挖掘的单位是账户、个人或家庭，还有其他一些单位。第 11 章有关于聚类城镇的案例研究，那是一家报纸开发编辑区域的行为准则，获取建模通常发生在地理区域层次、户口普查群体或邮政编码（zip code）层次。在客户关系管理之外的应用甚至更不相同。举例来说，

Mastering Data Mining 一书中有一个案例研究，其中的客户特征标识是杂志印刷厂中的出版物发行。

17.1.2 列

数据列包含描述客户某方面的数值。在有些情形中，列直接来自当前的商业系统；更常见的是，列是某些计算的结果，称之为衍生变量。

每列包含数值。范围指的是该列允许的取值集合。表 17-1 展示了数据挖掘使用的典型数据类型的范围特征。

表 17-1 数据挖掘使用的典型数据类型的范围特征

变量类型	典型范围特征
分类变量	可接受数值的列表
数值型	最小和最大值
日期型	最早和最晚日期，通常最晚日期小于或等于当前日期
货币金额	大于或等于 0
持续时间	大于或等于 0（或者严格地说大于 0）
分箱或分位数数值	分位数数字
计数	大于或等于 0（或者大于或等于 1）

直方图（histogram），如图 17-2 所示，显示每个数值或数值范围在某个数据集中出现的频率。纵轴是记录的计数，横轴是列中的数值。该直方图的形状表示数值分布（严格来说，在一个分布中，计数要除以记录总数，因此曲线下面的面积是 1）。如果使用随机选取的样本，那么在子集中的数值分布应该差不多与初始数据分布一样。

数值分布提供了对数据的重要深入了解。它表明哪些数值是常见的，哪些是比较罕见的。仅仅观察数值分布就引出一些问题，如数量为什么是负的，或为什么有些分类数值（categorical value）没有出现。虽然统计学家比数据挖掘者更关心分布，但观察变量值仍然很重要。此处，我们既列举了一些对数据挖掘目的相当重要的特殊分布案例，还列举了与目标一致的特殊变量案例。

1. 带有一个数值的列

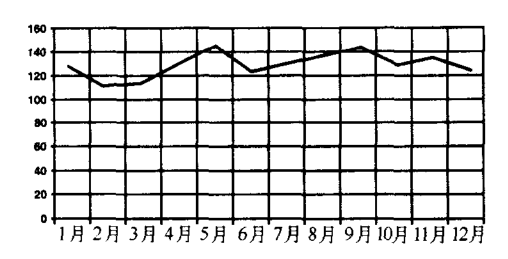
退化最严重的分布是只有一个数值的列。一元数值列，顾名思义，不包含任何可以帮助区分不同行的信息。因为缺乏任何信息内容，对于数据挖掘目的而言，它们应该被忽略。

只有一个值有时是数据的特性。举例来说，一个数据库拥有尚未开发的数据库中定义的字段并不罕见。字段是未来数值的惟一占位符，因此所有数值统一使用一个标识，如“null”或“no”或“0”。

在排除一元变量之前，检查 NULL 被当作数值的计数。附加的人口统计变量有时只有单一数值，或者当数值不为人所知的时候使用 NULL。例如，如果数据提供者知道某人对打高尔夫球感兴趣，因为他订购了高尔夫球杂志或加入了某个地区俱乐部，那么“高尔夫球迷”的标志就被设为“Y”。当没有证据时，许多数据提供者设定该标志为 NULL，这意味着不确定，而不是“N”。

提示：当变量只有惟一数值时，要确定：（1）NULL 被当作数值的计数；（2）当

选择行时，其他数值不会因为疏忽被遗漏。

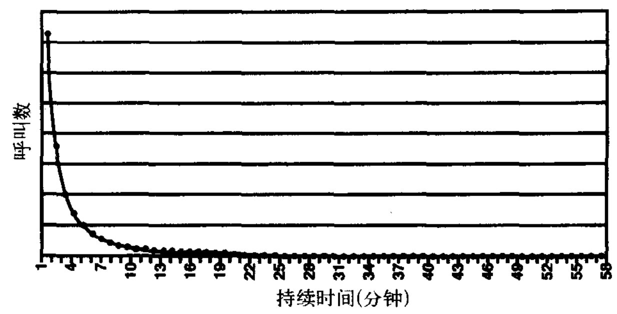


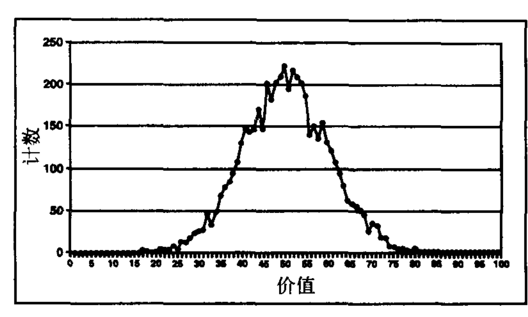
该直方图表示一组保险索赔的索赔月份

这是一个典型的均匀分布的例子。即每个月的索赔数大致相同

该直方图显示不同持续时间的呼叫电话数目

这是指数递减分布的例子





该直方图显示均值为50、标准差为10的正态分布。注意高值和低值非常少

图 17-2　直方图显示数据值的分布

当数据挖掘工作关注客户子集时，一元数值列也随之出现，用于过滤记录的字段在结果表保留下来。定义这个子集的字段可能都包含相同的数值。例如，如果在新泽西州为汽车客户构建模型，预测损失比率（一个保险度量标准），那么州的字段总是填写“NJ”字样。对所使用的样本，这个字段没有任何信息，因此，为了建模的目的，它应该被忽略。

2. 几乎只有惟一值的列

在“几乎一元”的列中，几乎所有记录在该列都有相同的数值。可能有几个离群值(outlier)，但是非常少。举例来说，零售数据可能汇总在每个部门中每位客户的所有购物。极少客户会从食品杂货商店的汽车部或百货公司的烟草部进行购买。因此，几乎所有的客户从这些部门的总购买量为 0 美元。

购物数据时常也是以“几乎一元”的形式出现。除了少许人之外，对所有人来说，像“收集瓷娃娃的人”或“在高尔夫球场上的费用量”等字段，值都是 NULL 或 0 美元。某些数据，例如调查数据，只是对一个非常小的客户集合可用。这都是数据倾斜的极端例子，如图 17-3 所示。

“几乎一元”列的很大问题是：“何时可以忽略它们?”为了证明忽略它们是正确的，数

值必须具有两个特点。第一，几乎所有记录必须有相同的数值。第二，必须仅有少量记录带有不同数值，构成数据的一个可忽略的部分。

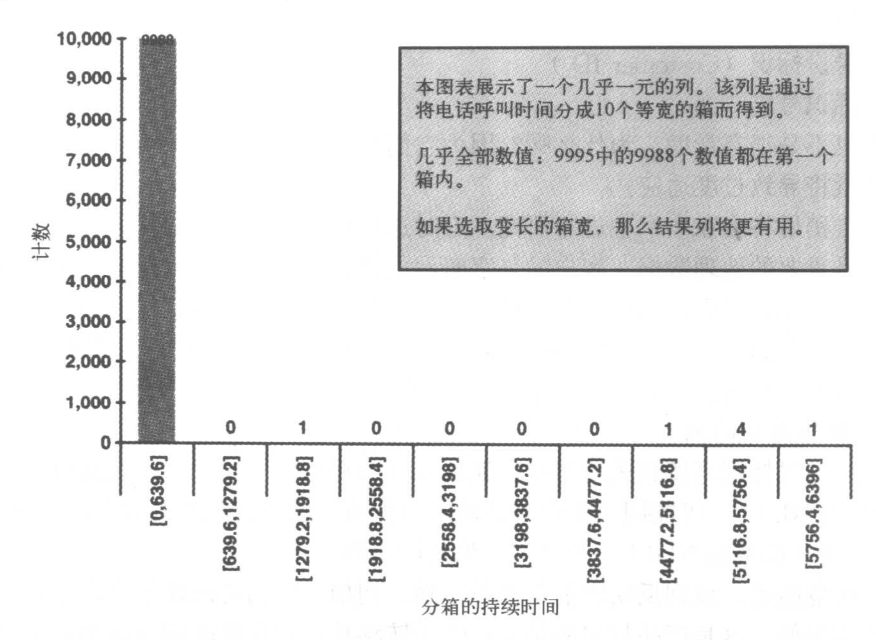


图 17-3　一个几乎一元的字段，如在本案例中由等宽分箱产生的箱，对数据挖掘目的毫无用处

数据的可忽略部分是什么？它是一个非常小的团体，即使数据挖掘算法完全能够识别，但由于团体太小，也显得不再重要。

然而，在忽略列之前，重要的是了解数值为什么如此严重倾斜。该列反映了关于商业的什么事情？或许极少数人曾经购买汽车产品，因为只有少数被调查商店曾经出售汽车。在这种情况下，按照“汽车产品－买主”标识客户，不可能是有用的。

在其他情况下，一个事件对于其他理由而言可能是稀罕的。在任何给定的一天，取消电话服务的人数可以忽略不计，但是随着时间的推移，数字日积月累。因此，需要在比较长的时期收集取消量，例如以月、季或年为单位。或者，收集精美玩偶的人数可能本来很少，但是当与其他字段结合的时候，可能就暗示一个重要的收藏家群体。

经验法则是，即使已经证明该列包含非常丰富的信息，但是如果它是几乎一元的，就不可能对数据挖掘有用。即，完全理解具有不同数值的行不能产生可操作的结果。作为一般的经验法则，如果列中 95%～99% 的数值相同，在孤立状态下，如果不进行一些处理，列很可能毫无用处。举例来说，如果令人质疑的列代表模型的目标变量，那么分层取样能产生一个样本集，其中的数据被高度集中。另外一种方法是结合几个这样的列，创建会被证明是很有价值的衍生变量。作为一个例子，某些户口普查地区的居民居住分散，例如那些特殊职业的地区。然而，将某些字段联合成单一的字段，如“地位显赫的职业”，能证明对于建模目的是有用的。

3. 带有惟一数值的列

对每个单一行或者几乎每一行取不同数值的分类列属于另一个极端。这些列惟一地（或者非常接近）识别每位客户，例如：

- 客户名字
- 地址
- 电话号码
- 客户身份标识（Customer ID）
- 车辆标识号码

这些列也不是很有帮助。为什么呢？因为它们惟一地识别每行，所以它们没有预言性价值。这种变量将导致过度适应。

一条忠告稍后将会在本章中进行研究。有时这些列包含很丰富的信息。在电话号码和地址中潜藏的是重要的地理数据。客户的名字暗示了性别。客户号码可能是按时序分派的，说明哪些客户是近期开通的，因此在决策树中揭示重要的变量。这些是从字段中提取重要特征（例如地理布局信息和客户崭新度）作为衍生变量的案例。然而，数据挖掘算法还远未强大到提取来自数值的这种信息，需要数据挖掘者进行提取。

4. 与目标相关联的列

当某一列与目标列高度相关的时候，就意味着列只是一个同义字。在这里举两个例子：

- “账号为 NULL”可能同义于营销活动响应失败。响应者只是开立账户并被分配账号。
- “流失的日期不是 NULL”与已经流失是同义的。

另外一种危险是，该列反映以前的商业实践。例如，数据可能显示具有呼叫转移的所有客户也有呼叫等候，这是产品打包的结果；呼叫转移总是在包括呼叫等候的打包产品中被卖出。或者数据可能显示，几乎所有的客户居住在最富有的地区，因为这里是过去的客户获取活动的目标。该例说明，数据挖掘者需要了解历史商业实践。与目标同义的列应该被忽略。

提示：一种容易找到与目标同义的列的方法是建立决策树。决策树将会选择一个同义变量，然后这个变量可以被忽略。如果决策树工具让你见到其他可能的拆分，那么能立刻发现所有这类变量。

17.1.3 模型在建模中的角色

列包含带有数据类型的数据。除此之外，列具有数据挖掘算法相关的角色。三个重要的角色是：

1）**输入列**。即那些用做模型输入的列。

2）**目标列**。即仅用于构建预言性模型的一个列或者一组列。这是一些值得关注的事情，如购买特别产品的倾向性（propensity）、响应优惠的概率或者保留客户的可能性。当构建非定向模型时，就不需要有目标。

3）**已忽略的列**。即不再使用的列。

不同的工具中，这些角色有不同的名字。图 17-4 展示了在 Angoss Knowledge Studio 中如何去除一个列。

提示：被忽略的列在聚类中起着非常重要的作用。由于被忽略的列不能用来建立簇，它们在簇中的分布可能很有价值。通过忽略如客户利润率或响应标志等列，能够发现这些被忽略的列是如何在簇中分布，还可能正好发现了关于客户利润或响应者的非常关键的事情。

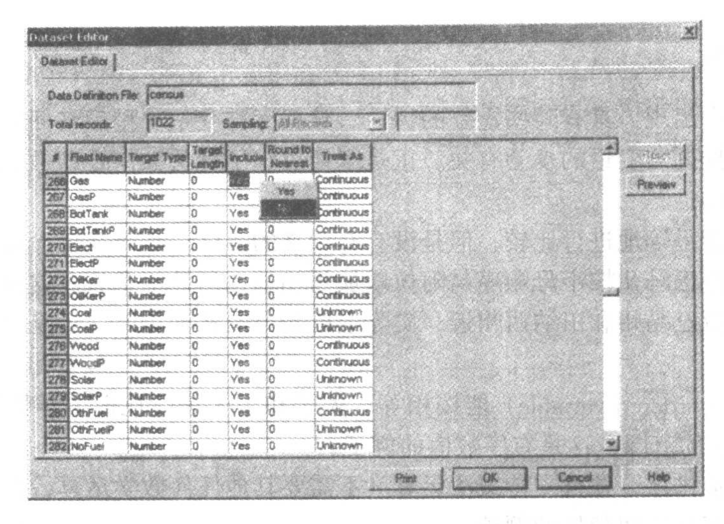


图 17-4 Angoss Knowledge Studio 支持几个模型角色，例如构建模型时忽略一个列

也有某些更高级的角色能够在特定的环境之下使用。图 17-5 显示了在 SAS Enterprise Miner 中许多可用的模型角色。这些模型角色包括：

1) **标识列**。是惟一识别每行的列。总而言之，这些列对于数据挖掘目的可以忽略，但是对于评分很重要。

2) **权重列**。详细说明适用于每行的“权重”。是通过包含数据权重创造权重样本的方法。

3) **成本列**。详细说明与行相关联的成本。举例来说，如果正在构建保留客户模型，那么“成本”可能包括每位客户价值的估计。有些工具能够使用这种信息优化正在构建的模型。工具中另外的模型角色是 SAS Enterprise Miner 所特有的。

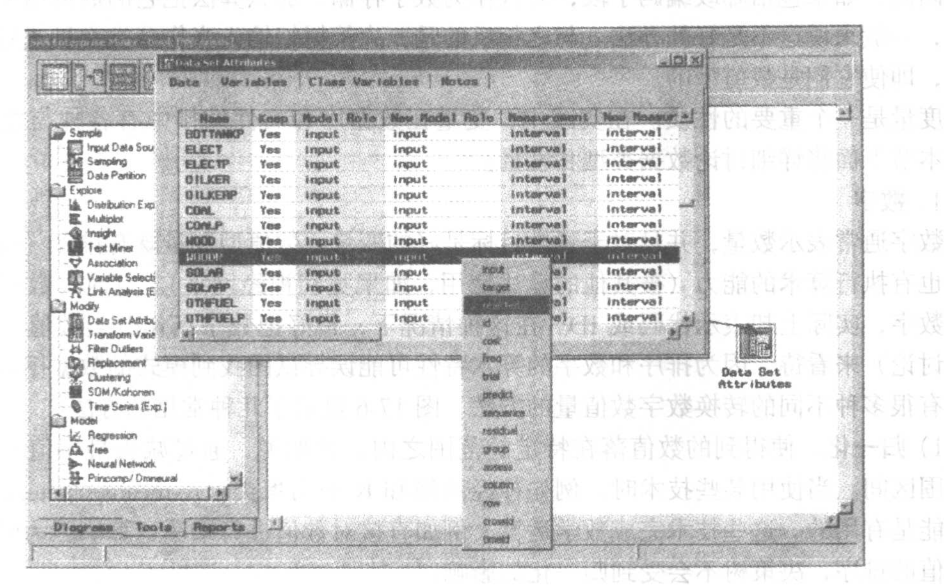


图 17-5 SAS Enterprise Miner 具有很宽范围的可用模型角色

17.1.4 变量度量

变量在数据中出现，具有某些重要的特性。数据库与变量（稍后将回到这个主题）的类型有关，而数据挖掘与变量的度量有关。正是度量决定了算法如何处理数值。下面的度量对数据挖掘是重要的：

- *分类变量*能平等地进行比较，但是没有有意义的排序。举例来说，州的缩写属于分类变量。阿拉巴马州按字母顺序与阿拉斯加州相邻，但并不意味着这两个州之间的实际距离比阿拉巴马州和田纳西州近，实际上后两个州相互接壤，但是要按字母顺序就显得非常远。
- *有序变量*（ordered variable）能按相等、大于和小于进行比较。学生课堂成绩可以归入到 A～F 的范围，这是有序数值的例子。
- *区间变量*被排序，并且支持减法运算（不需要其他任何数学运算，例如加法和乘法）。日期和温度是区间变量的例子。
- *真数值型变量*（true numeric variable）是支持加法和其他数学运算的区间变量。钱数和客户保有期（按数天计算）是数值型变量的例子。

真数值和区间之间的区别是细微的。然而，数据挖掘算法以相同的方法对待它们。同时要注意，这些度量形成分层：任何有序变量也是分类变量，任何区间变量也是分类变量，任何数值型变量也都是区间变量。

度量和数据类型之间有区别。例如，数值型变量可能表示编码方案，如表示账号状态或甚至州的缩写。虽然数值看起来像数字，实际上属于分类。邮政编码是这种现象的普通例子。

某些算法期望变量具有某种度量。举例来说，统计回归和神经网络期望输入是数值型的。因此，如果包括邮政编码字段，并且作为数字存储，那么算法把它的数值当作数值型来看待，一般来说这不是好的办法。与之相反的是，决策树把输入当作分类变量或有序变量来看待，即使它们是数值型的。

度量是一个重要的性质。在实践中，变量在数据库和文件编排中有各种与之关联的类型。本节下面将详细讨论数据类型和度量。

1. 数字

数字通常表示数量，并且对于建模目标是好的变量。数值型数量既有排序（被决策树使用）也有执行算术的能力（被其他的算法使用，如聚类和神经网络）。有时，数字看起来像一个数字，实际上却表示代码或 ID。在这种情况下，最好把数字当作分类数值（在以下两部分讨论）来看待，因为排序和数字的算术特性可能误导试图找到模式的数据挖掘算法。

有很多种不同的转换数字数值量的方法。图 17-6 显示了几种常见的方法：

1）**归一化**。使得到的数值落在特定的范围之内，比如说，通过减去最小值并且除以整个范围区间。当使用某些技术时，例如神经网络和 K 平均聚类（K-means clustering），归一化可能是有用的。这些技术完成数学运算，例如直接对数值进行乘法运算。因为归一化不改变数值的排序，决策树不会受到归一化的影响。

2）**标准化**。即把数值转变成偏离均值的标准差数量，它很好地揭示了数值的非经期程度。用到的算法很容易——减去平均值并且除以标准差。这些标准值也被称为 z 得分。和归

一化一样，标准化不影响排序，因此，它对决策树没有影响。

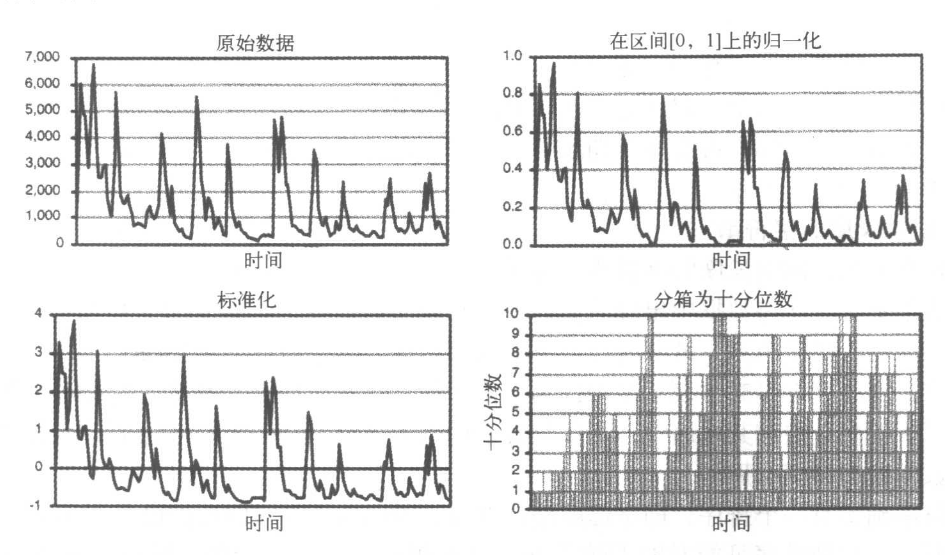


图 17-6 归一化、标准化和分箱是转换数值型变量的典型方法

3) **等宽分箱**。即把变量转变成固定宽度的范围。形成的变量和初始变量有大致相同的分布。然而，分箱（binning）数值影响所有的数据挖掘算法。

4) **等高分箱**。即把变量转变成 n 个块（例如五分位数或十分位数），因此相同数目的记录落入每个箱子。得到的变量呈均匀分布。

或许料想不到的是，分箱数值能改善数据挖掘算法的性能。对于神经网络来说，分箱是减少离群值影响力的几种方法之一，因为所有的离群值进入同一个箱内。对于决策树来说，分箱变量可能造成在树的高度层次上有较多相等大小的子结点（即，与出现一个子结点得到5%的记录而其他子结点得到95%的记录的情况相反，利用相应的分箱变量，一个箱子可能得到20%，而其他箱子得到80%）。虽然基于分箱变量的拆分不是最优结果，但是此后的拆分可能产生更好的树。

2. 日期和时间

日期和时间是区间变量最常见的例子。因为这些变量将时间元素引入到数据分析，所以它们非常重要。通常，日期和时间变量的重要意义在于，它们为其他的变量提供时间序列和时间戳信息，例如最后一次投诉呼叫的原因和分析。

因为有无数不同的格式，采用日期和时间戳的工作可能是困难的。Excel 有 15 种不同的为单元预置的日期格式，并且具有定制更多形式的能力。日期和时间的一种典型的内在格式是作为单个数字，即从过去某个日期开始的天数或秒数。当这种情形出现时，数据挖掘算法把日期当作数字看待。这种表示足以使算法检测较早和稍后发生的事情。然而，也错过其他值得加入数据的重要特性：

- 一天某时
- 一周的某一天，是否是工作日或者周末
- 假日

Ralph Kimball 在他的 *The Data Warehouse Toolkit*（Wiley，2002）一书中大力推荐日历

是为数据仓库创建的首选表格之一。由于日历的属性通常对数据挖掘工作很重要，我们也非常赞同这个建议。

当采用日期和时间工作时，遇到的一个困难就是时区。尤其是在互相连接的网络世界中，时间戳通常是来自服务器计算机的时间戳，而不是客户所在位置的时间。记住，在凌晨访问网站的客户，实际上可能是新加坡的一位在午餐时间上网冲浪的人，而不是生活在纽约的夜猫子。

3. 固定长度的字符串

通常意义上，固定长度的字符串表示分类变量（categorical variable），其取值属于一个已知的数值集。应该将数据中出现的实际数值与合法数值列表进行比较，以检查非法数值，查证该字段是否已被填充，考察哪些数值是最频繁的，哪些是最不频繁的。

固定长度的字符串时常表示某种代码。最好是，时常有参照表描述这些代码的含义是什么。参照表提供分层和其他属性，当仅仅观察代码本身时，这些属性可能是不明显的，所以它们对数据挖掘特别有用。

字符串确实有一个排序，即字母顺序。然而，如同前面列举的阿拉巴马州和阿拉斯加州的例子所示，这种排序可能对图书馆管理员是有用的，但是对数据挖掘者来说用处不是很大。当存在可判断的次序时，用数字代替代码是有意义的。举例来说，一家公司将客户划分成三个群体：少于 1 年保有期的“新”客户、在 1 和 2 年之间的“边缘”客户和超过 2 年的“核心”客户。这些范畴具有清晰的排序。实际上，具体化的排序的一种方法可能是将三个群体分别映射到数字 1、2 和 3。更好的方法是要包括用于数据挖掘目的的真实保有期，尽管报告可能仍然以保有期分组为基础。

当有较少的分类时，数据挖掘算法通常表现更好。减少大量分类的一个方法就是使用代码属性，而不是代码本身。举例来说，一家移动电话公司的客户可能使用数百种不同的移动电话设备代码（尽管只是几种流行的产品吸引巨量客户）。我们不是独立地使用每个型号，而是包括移动电话的重量、移动电话最初上市的日期等特征，以及由此提供的一些特征。

美国的邮政编码提供了取值很多、潜在有用的变量的好例子。减少数值数目的一种方法是只使用前三个字符（数字），这些是区域中心设施（SCF），通常处于县或大城镇的中心。它们保持邮政编码中大部分的地理信息，但处于更高层次。尽管 SCF 和邮政编码都是数字，但应该都视作代码。需要提及的是，在邮政编码中，开头的数字“0”很重要，例如 Data Miners 公司的邮政编码是 02114，如果没有开头的数字“0”，就没有任何意义。

有些业务是区域性的，结果是，几乎所有的客户都位于少数几个邮政编码区域。然而，仍然可能有很多其他的客户松散地分布在许多其他的地方。在这种情况下，最好是把所有松散数值分到单个“其他”类。另外一个较好的方法是，用关于邮政编码的信息代替邮政编码。可能有几条信息，例如中值收入和平均住宅价格（来自人口普查局），还有最近一次营销活动的穿透度和响应率。用描述性数字更换字符串数值是将商业知识引入建模的有力方法。

提示： 用分类的数值型汇总（例如在一个邮政编码区域内的产品穿透度）代替分类变量，能够改善数据挖掘模型，解决分类中有很多数值的问题。

神经网络和 K 平均聚类是期望输入是区间变量或真数值型变量的算法的例子。这就引发了关于字符串的问题。幼稚的方法是为每个数值分配一个数字。然而，数字含有代码中没

有的信息，例如排序。这种虚假的排序能隐藏数据中的信息。更好的方法是给每一个可能的数值创造一组标志，称为指示器变量（indicator variable）。虽然这增加了变量的数量，但消除了虚假排序的问题，改善了结果。神经网络工具时常自动进行这些工作。

总之，有几种方法处理固定长度的字符串：

- 如果仅有几个数值，那么直接使用数值。
- 如果数值包含有用的排序，那么数值能被变成代表排序的分级。
- 如果有参照表，那么描述代码的信息可能是更有用的。
- 如果几个数值占有主导优势，但也有很多其他数值，那么较稀有的数值可以分到一个“其他”类中。
- 对于期望数值型输入的神经网络和其他算法，数值可以映射到指示器变量。

这些方法的共同特征是，将域信息纳入编码程序，因此，数据挖掘算法能够寻找料想不到的模式，而不是发现已知的模式。

4. 身份标识（ID）和关键字

一些变量的目的是提供到有较多信息的其他记录的链接。身份标识和关键字时常被作为数字加以存储，尽管也可能以字符串的形式存储。作为一般的规则，这种身份标识和关键字不应该直接用于建模目的。

数据挖掘应该忽略的字段的一个好例子是账号。出人意料的是，此类字段可能改善模型，因为账号不是随机分配的。时常，它们被顺序分配，因此旧账号的号码低；也可能基于获取渠道分配，因此所有网络账号比其他账号的号码高。最好在客户特征标识中明确包括有关的信息，而不要依赖于潜在的商业规则。

在有些情形中，身份标识确实加入了有意义的信息。在这些情况下，应该提取信息，使它更接近数据挖掘算法的需要。下面给出一些例子。

*电话号码*包含国家代码、区号和电话交换局，所有这些都包含地理信息。在北美地区，标准的10位数电话号码前三位代表区号，后三位表示电话局，最后四位代表电话线路。在大多数数据库中，区号提供有益的地理信息。在北美地区以外，电话号码的格式各有不同。在某些情形中，区号和电话号码是变长的字符串，使提取地理信息更加困难。

统一产品代码（A类UPC）是12位代码，识别通过扫描仪的许多产品。前六位是制造商代码，紧跟着的五位代表特定产品的代码。最后一位数字没有具体意义，是用来检验数据的校验数字。

*车辆标识码*是刻在汽车上的17个字符的代码，描述制造商、型号和车辆的生产年。第一个字符描述原产国家；第二个代表制造商厂家；第三个是车辆类型；第4～8个字符记录车辆的特定特征；第10个是此款车辆的生产年；第11个是生产车辆的装配厂；剩余的六个是连续的产品序列号。

*信用卡号*有13～16位数字。前几位数字是卡片网络代码。特别是，它们能区别美国特快卡（American Express）、维萨卡（Visa）、万事达卡（MasterCard）和发现号卡（Discover），等等。不幸的是，其他数字的使用依赖于网络，因此，没有统一的标准来区分金卡和银卡。顺便提一下，最后一个数字是用来作为基本的校验数字，用来验证信用卡号是否有效。校验数字的算法被称为Luhn算法，以IBM公司开发它的研究人员命名。

在一些国家（不是美国），*公民身份标识码*含有个人性别和出生的数据。当它可用的时

候，这是有益的、精确的人口统计信息资源。

5. **名字**

虽然我们想了解客户，但数据挖掘的目标不是真正面对面接触他们。一般而言，名字对于数据挖掘不是有用的信息源。当试图了解特定的市场，或按性别发送信息时，有一些情形，依照种族（像西班牙名字或亚洲人的名字）分类名字可能是有意义的。然而，这种工作充其量是非常粗糙的近似，并且不会被广泛应用于建模目的。

6. **地址**

地址描述客户的地理信息，对了解客户行为非常重要。不幸的是，只有邮局能够理解许多不同的书写地址方式的变形。幸运的是，有服务局和软件能够标准化地址字段。

地址最重要的用途之一，是了解两个地址什么时候指的是同一地址，什么时候是不同的。举例来说，在网络上订购产品的递送地址是否与银行信用卡的账单地址相同？如果不是，可能暗示购买的是一件礼品（如果两个地址相距很远，且支付了包装礼物的费用，则这种暗示更强烈）。

除了发现精确匹配之外，整个地址本身不是特别地有用；最好提取有用的信息，用另外的字段表示它。某些有用的特征如下：

- 公寓号（有或没有）
- 城市
- 州
- 邮政编码

最后三个通常被存储在不同的字段中。因为地理学时常在理解客户行为方面起着比较重要的作用，所以我们推荐标准化地址字段，并附加有用的信息，如户口普查群组、多单元楼或单个单元楼、居住地址或商务地址、纬度、经度，等等。

7. **自由文本**

自由文本向数据挖掘提出挑战，因为这些字段提供丰富的信息，通常很容易被人类理解，但是不能被自动化的算法领悟。已经发现，最佳方式是从文本中巧妙地提取特征，而不是向计算机展现整个文本字段。

文本有许多来源，例如：

- 医生诊视病人的记录
- 呼叫中心人员打印的备忘录
- 发送客户服务中心的电子邮件
- 以表格形式提交的评论，不管是网络表格还是保险表格
- 在呼叫中心的声音识别算法

在商业界的文本源具有特定的特性，它们不合乎文法，并且充满了错误的拼写和缩写。人类一般能够理解它们，但是对于自动化这种理解是非常困难的。因此，即使人们容易辨识多余的邮件，编写自动过滤垃圾邮件的软件也是相当困难。

我们推荐的方法是，通过寻找特定的子串探索特别的特征。举例来说，从前一个犹太人群体因为一家公司支持以色列的立场而联合抵制这家公司。呼叫中心服务员打印的备忘录字段是关于为什么客户停止的最佳信息来源。不幸的是，这些字段不是统一地表示“由于以色列的政治原因而停止”。事实上，许多评论包含了一些对“Isreal”（以色列）、“Is rael”、

“Palistine” 等的引用（编者注：Palestine，巴勒斯坦）。分类文本备忘录需要在文本（在这种情况下，“Israel”、“Isreal” 和 “Is rael” 都被使用）中寻找特定的特征，然后分析结果。

8. 二进制数据（声音、图像等）

不用惊奇，有其他一些数据类型没有落入这些很好的类。声音和图像变得日益普遍，但数据挖掘工具通常不支持它们。

由于这些类型的数据可能包含丰富的数据，可以对它们做什么呢？答案是提取特征放入衍生变量之内。然而，这种特征提取工作对所使用的数据是非常特别的，并且已超出本书的范围。

17.1.5 用于数据挖掘的数据

数据挖掘期盼数据有特别的格式：

- 所有数据应该放在单一表格中
- 每行应该与一个实体相对应，例如客户，与商务有关
- 带有单一数值的列应该被忽略
- 对每列带有不同数值的列应该被忽略，虽然它们的信息可能被包含在导出列之中
- 对于预言性建模，目标列应该被识别，并且所有的同义列要除去

唉，这不是在现实世界中发现数据的方式！在现实世界中，数据来自于源系统（source system），可以用特别的方式存储每个字段。通常，我们需要使用存储在参照表中的数值代替字段，或者从更复杂的数据类型提取特征。下一节讨论把这些数据整理成为客户特征标识。

17.2 构建客户特征标识

构建客户特征标识，尤其第一次，是一个逐渐递增的过程。最低要求，客户特征标识至少需要构建两次，其中一次构建模型，一次用于评分。实际上，探索数据和建立模型提出新的变量和转换，因此，需要多次重复这个过程。具有可重复的过程使数据挖掘工作变得简单。

如图 17-7 所示，过程中的第一步是识别数据的有效来源。毕竟，就客户层次而言，客户特征标识是概要，是已知的关于客户的信息。概要以可用的数据为基础，这笔数据可能存在于数据仓库中，也可能存在于操作系统中，有一些可能是由外部厂商提供。当进行预言性建模的时候，识别目标变量的来源特别重要。

第二个步骤是识别客户。在某些情形中，客户停留在账户层次。在其他情形中，客户处于个体或家庭层次。在某些情形中，客户特征标识可能与某个人一点关系也没有。举例来说，我们已经使用客户特征标识来了解产品、邮政编码和县，尽管客户特征标识最普通的用途是账户和家庭。

一旦客户被识别，数据来源需要被映射到客户层次。这可能需要另外的查找表（lookup table），例如，把账户转换到家庭。在已有的数据中发现客户是不可能的。在这种情形下，需要再次访问客户定义。

构建客户特征标识的关键是从简单开始，并且逐步发展。按照将数据源映射到客户的难易程度，对它们进行优先排序。从最容易的一个开始，并且用它建立客户特征标识。在加入

所有数据之前，也可以使用客户特征标识。当等待比较复杂的数据转换（data transformation）时，开始做并且理解什么是可用的。当从交易中构建客户特征标识时，确保得到与特定客户相关联的全部交易。

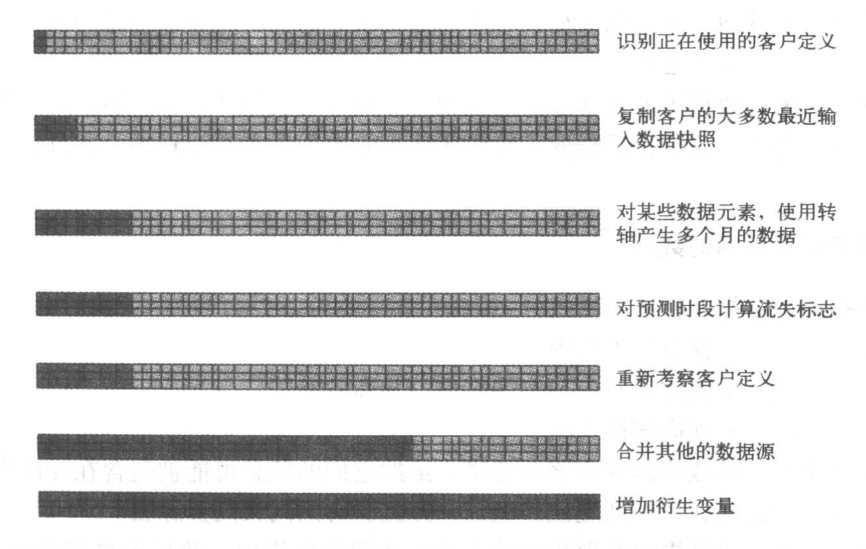


图 17-7　构建客户特征标识是一个反复的过程；按部就班地从小的开始，贯穿整个过程，如同在本例中为流失预测构建客户特征标识一样

17.2.1　编写数据目录

在移动通信公司，数据挖掘工作组想要开发内部的流失模型。这个流失模型给定一个月的延迟时间来预测一个月的流失。因此，如果二月的数据可用，那么流失预测的是四月份。这种模型为收集数据以及给新客户评分提供时间，因为有时二月的数据在三月的某个时间才能得到。

在这家公司，客户特征标识有几个潜在的数据来源。18 个月的历史数据全部保留在数据存储库中。基本上，每个文件是月末结束时操作系统转储到数据存储库的快照。

UNIT_MASTER 文件包含服务中每个电话号码的描述，以及在月末时了解的电话号码的快照。在这个文件中，作为字段的例子是电话号码、账单账户、电话套餐、移动电话型号、最后发送账单日期和最后付款。

TRANS_MASTER 文件包含在每个月期间发生在特定电话号码的每笔交易。这些是账户层次的交易，包括连接、切断、移动电话升级等。

BILL_MASTER 文件在账户层次描述账单信息。多个移动电话可能被附加到相同的账单账户上，特别是那些商业客户和使用家庭电话套餐的客户。

虽然其他的数据来源在这家公司是可用的，但是不会立刻突出用于客户特征标识。举例来说，一个来源是呼叫的详细记录，即每个电话呼叫的记录，对预测流失是有用的。虽然这笔数据最终被数据挖掘工作组使用，但却不是最初工作的组成部分。

17.2.2　识别客户

数据是现实世界的典型代表。虽然数据关注点可能在某一类客户上，但数据有多个群

体。后面“居民客户和商务客户”部分谈论这两者之间的区别。

在这个例子中谈到的商业问题是流失。如图 17-8 所示，客户数据模型相当复杂，导致客户定义有不同选择：

- 电话号码
- 客户身份标识（ID）
- 账单账户

然而，这就是真实世界，重要的是记住这些关系是复杂的，并且随时间变化。客户可能变换电话号码，电话可能被加到账户或从账户中删除，客户可能改变移动电话，等等。为了构建客户特征标识，决策是使用电话号码，因为这正是企业报告流失的手段。

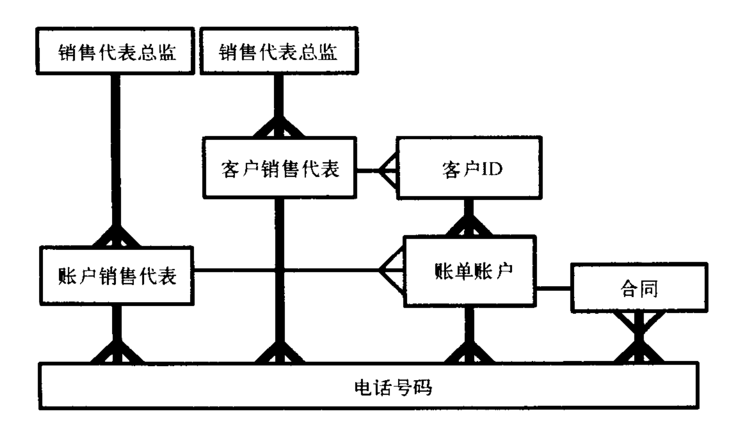


图 17-8 客户模型是复杂的，并且考虑销售、账单和业务分层信息

居民客户和商务客户

通常，数据挖掘工作集中于一种客户类型，例如居民客户或小企业。然而，所有客户的数据时常在操作系统和数据仓库中混杂在一起。通常，有很多方法区别这些客户的类型：

- 时常有客户类型字段，取值像“居民”和“小企业”。
- 可能有一个销售分层；某些销售渠道是商务专用，而其他一些是居民专用。
- 有些电话套餐只适用于企业；而有些只适用于居民客户。
- 可能有一些企业规则，因此超过两条线的任何客户被看做企业。

这些例子阐述这样一个事实，有几种不同的典型规则可以区分不同的客户类。假设机会是不一致的，大多数的数据来源不会无效。不同的规则选择不同的客户子集。

这是问题吗？那要依赖所工作的特定模型。希望的是规则都非常接近，因此，根据一条规则包括进来（或错过）的客户本质上与根据其余规则包括的客户是相同的。重要的是调查这是否是真实地，以及何时规则是不一致的。

在实践中常常发生的是规则之一居支配地位，因为这就是企业的组织方式。因此，客户类型可能是重要的，销售分层或许更重要，因为这与不同的客户片段负责人相对应。

在企业和居民之间的差别对于潜在客户和客户同等重要。一家长途电话公司看到网络上的很多呼叫是由其他电信公司的客户呼出。交换机生成了呼叫明细记录，包括呼出号码和目标号码。任何不属于现有客户的家庭号码就是一位潜在客户。一家长途电话公司建立客户特征标识，来描述未知电话号码随时间变化的行为，追踪诸如该号码出现的频繁程度，在一天

的某个时间或者一周的哪些天是相当活跃，以及典型的呼叫持续时间。此外，这种客户特征标识可用来获得未知电话号码是企业客户的可能性，因为企业客户和居民客户被不同的优惠服务所吸引。

如果目的是为居民客户构建模型，一种简化是，仅仅关注只有一个电话号码的客户账号，这是开始简化数据模型的一种好方法。如果目标是为企业客户构建模型，更好的客户层次是账单账户层次，因为企业客户时常将移动电话和电话号码打开和关闭。然而，在这种情况下，流失意味着取消整个账户，而不是单个电话号码。对于那些只有一条电话线路的居民客户来说，这两种情形是相同的。

17.2.3 第一次尝试

为了建立客户特征标识，第一次尝试需要集中在最简单的数据源。在这种情况下，最简单的数据来源是 UNIT _ MASTER 文件，它在电话号码层次上方便地存储数据，这个层次正是客户特征标识所用的层次。

值得指出的是，这个文件和客户定义存在两个问题：

- 客户可能改变电话号码
- 电话号码可能被重新分配给新客户

这些问题将会在稍后的部分讨论；第一个客户特征标识是在电话号码层次上开始。用来构建客户特征标识的过程分为四步：识别时间帧（time frame），创建最近的快照，转轴（pivoting）列和计算目标。

1. 识别时间帧

在构建客户特征标识时，第一次尝试需要考虑数据的时间帧，正如第 3 章中的讨论。图 17-9 显示这笔数据的模拟时间图表。最终的模型集中应该至少包含一个以上的时间帧。然而，第一次尝试只关注一个时间帧。

	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月
得分						4	3	2	1		P
模型集				4	3	2	1		P		
模型集			4	3	2	1		P			

图 17-9 一个模拟时间图表，展示了创建客户特征标识时的输入列和目标的时间帧

时间帧定义了在 8 月份一个月期间的流失。所有输入数据都至少来自一个月以前。为了提供一个月的等待时间，截止日期是 6 月 30 日。

2. 获得最近的快照

数据的最近快照按照截止日期定义。客户特征标识中的这些字段描述了在客户流失（或没有流失）之前，已知的关于客户的最近信息。

这是来自 6 月份的 UNIT _ MASTER 文件的一组字段，如移动电话类型、电话套餐，等等。当填充客户特征标识的时候，考虑时间帧是重要的。为避免混乱，可以使用命名规则。在本案例中，所有的字段可能都有一个后缀“_ 01”，表明它们来自最近一个月的输入数据。

提示：当构建客户特征标识的时候，使用命名规则表明每个变量的时间帧。举例来说，

最近一个月的输入数据可能有“_ 01”后缀；在此之前的一个月，后缀为“-02”；依此类推。

此时，关于该字段知之甚少，因此描述性的信息是有用的。举例来说，电话套餐可能有一个描述，如每月的基数、每分钟的费用，等等。所有这些特征是值得关注的，并且对建模有潜在价值，因此，把它们引入模型集是合理的。虽然描述不会用于建模（代码更好些），但是能够帮助数据挖掘者理解数据。

3. 转轴列

在 UNIT _ MASTER 文件中，一些字段代表在正则时间序列（regular time series）中报告的数据。举例来说，在账单数量中每个月有一个值，并且每个值需要放入单独的列中。这些列来自 UNIT _ MASTER 的不同记录，一列代表 6 月，一列代表 5 月，一列代表 4 月，等等。例如，使用命名规则，字段将会是：

- Last _ billed _ amount _ 01 代表 6 月（可能已经在快照内）
- Last _ billed _ amount _ 02 代表 5 月
- Lastbilled _ amount _ 03 代表 4 月

此时，客户特征标识开始形成。虽然输入字段只是来自一个来源，适当的字段已经适时地被选为输入并且被及时排列。

4. 计算目标

对于预言性建模，客户特征标识没有目标变量是不可能有用的。既然客户特征标识将被用于流失模型，目标需要是客户是否在 8 月份流失。在 8 月的 UNIT _ MASTER 记录中，这是账户状态字段。注意，只有在 6 月 30 日或之前活跃的客户被包含在模型集中；不包含 7 月开始、8 月取消的客户。

17.2.4 取得进展

客户特征标识尽管相当不完善，但现在已经可以在模型集中使用。由于有明确定义的时间帧、目标变量和输入变量，它是实用的，至少最低程度是这样。虽然客户特征标识是有用的，是良好的起点，但遗漏了几件事情。

首先，客户定义没有考虑电话号码的变化。因为 TRANS _ MASTER 文件追踪客户账户的变化类型，所以它解决了这个问题。为修复客户的定义，需要创建一个表格，包含账户的最初电话号码（或许带有一个计数器，因为电话号码可能实际上被重复使用）。在这张表格中，一个典型的行会有以下列：

- 电话号码
- 有效日期
- 结束日期
- 惟一的客户标识符

利用这张表格，客户标识符能代替电话号码使用，因此，客户特征标识对电话号码的变化明察秋毫。

客户特征标识的另一个缺点是它只依赖于一个数据源。应该增加另外的数据源，每次增加一个，以建立客户行为的更丰富的客户特征标识。模型集只有数据的一个时间帧，更多的时间帧可以使模型更稳定。这个客户特征标识也缺乏衍生变量，它是本章其余许多部分讨论

的主题。

17.2.5 实际的问题

当构建客户特征标识的时候，会遇到一些实际的问题。客户特征标识时常把最大的数据源合在一起，并且在其上进行复杂的操作。这在计算资源方面成为一个问题。虽然结果模型集可能至多有数十或数百兆字节，但是被汇总的数据可能是数千倍之大。因此，最好在关系数据库中尽可能地多做处理，因为这些操作能同时利用多个处理器和若干磁盘。

虽然最后得到的查询较复杂，但是整合客户特征的多数工作可以用 SQL 或数据库脚本语言进行。这是有用的，不仅因为它提高效率，而且因为代码只存储在一个地方，即减少错误的可能性，以及发现缺陷（bug）的能力。二者择其一，数据可以从源中抽取，然后拼凑起来。逐渐地，数据挖掘工具能够更好地利用数据。然而，这通常需要一定数量的编程，例如，使用编程语言 SAS、SPSS、S-Plus 或者 Perl。附加处理不仅增加工作的时间，而且引出第二个层次，在这个层次，缺陷可能会悄悄混入。

当创建客户特征标识的时候，意识到数据挖掘是一个反复、时常需要重建客户特征标识的过程是重要的。一个好的方法是，为从数据源抽取数据的一个时间帧建立模板，然后多次进行抽取，产生模型集。对于评分集，可以应用同样的程序，因为评分集与模型集非常类似。

17.3 探查变量

数据探查（data exploration）与数据挖掘过程高度相关。在许多情形下，数据挖掘和数据探查是实现共同目标而又相互补充的方法。数据挖掘倾向于突出发现模式的有意义的算法，而数据探查更关注表现数据，从而使人们能够凭直觉获知模式。当交流结果的时候，显示正在发生事情的精美图片时常比单调乏味的数字表格更有效。类似地，当为数据挖掘准备数据的时候，查看数据可以提供正在发生的事情的深入了解，这种深入了解有助于改进模型。

17.3.1 直方图分布

当查看数据的时候，开始的地方是每个域的直方图；直方图展示了域中数值的分布。实际上，因为直方图计算出现次数，而分布是归一化的，所以在直方图和分布之间有细微的不同。可是，就我们的目标而言，相似性更加重要，直方图和分布（或者严格地来说，与分布相关联的密度函数）有相似的形状，只是 Y 轴的标度有变化。

大多数数据挖掘工具提供将单一变量的值呈现为直方图的能力。纵轴表示每个数值在样本中出现的次数，横轴表示各种不同的数值。

当创建直方图时，数值型变量时常被分箱。为了探查变量，这些箱子应该是等宽而不等高的。需要记住的是，等高分箱产生的箱包含相同的数值个数。包含相似记录数的箱对建模是有用的，然而，对理解变量本身没有太大的用处。

17.3.2 随时间变化

当时间元素注入到直方图的时候，也许最具有启迪作用的信息开始显露。在这种情况

下，只有单个变量的一个数值被应用。图表显示，这个数值出现的频率如何随时间改变。

作为例子，图 17-10 中的图表十分清楚地展示了关于数值“DN”在三月期间发生的事情。这类模式是很重要的。在这种情况下，当两个不同的系统被合并的时候，“DN”表示需要消除的重复账号。事实上，只有在见到这种模式和询问在这期间发生的事情的问题之后，才困难地做出这个解释。

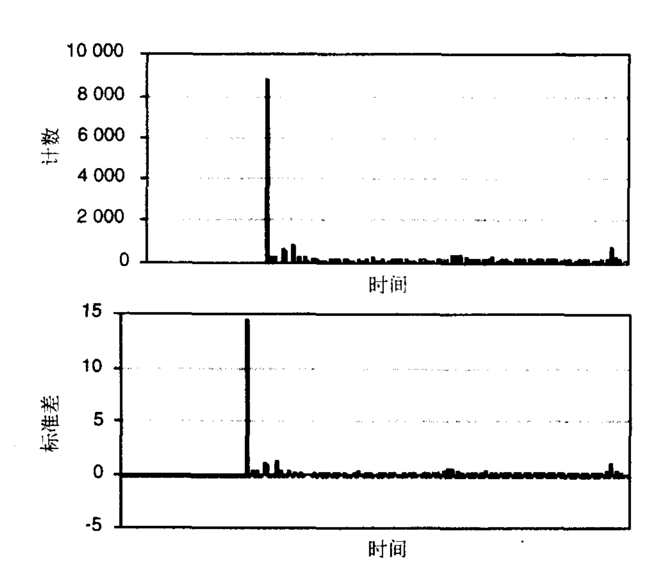


图 17-10　这个直方图意味着对于这个停止代码，不寻常的事情正在发生。在顶部的图表中是原始数据，而下面的图表中是标准化的数值

图表的顶部表示原值，并且可能是相当有用的。底部表示标准化的数值。在两个图表中，曲线的形状也相同；惟一的不同是垂直方向的标度。需要记住的是，标准化数值把它们转换成偏离均值的标准差，因此在 - 2 到 2 范围外的数值是不寻常的；小于 - 3 或者大于 3 的值应该是非常稀少的。同一数据的可视化显示，顶峰变出期望值许多个标准差——14 个标准差非常令人费解。这种随机发生的似然性是如此遥远，以致于图表暗示某种外部的事情正在影响变量，如以前两个计算机系统的合并等外部事件，如何创建了重复的账号。

按照时间创建一个交叉表并不困难。然而不幸的是，在数据挖掘工具中，对于这类图表没有很多支持。在 Excel 中，或在 SAS、SPSS、S-Plus 中，或几乎任何其他程序语言中，很容易用几行程序产生这类图表。问题就是需要许多这种图表，如对每一分类变量的每个取值都需要一个图表。例如下列情况：

- 按时间开设的不同类型账户。
- 客户按时间停止的不同理由。
- 某种地理布局随时间的特征。
- 不同渠道随时间的特征。

这些图表及时清晰地展现过去，它们引出何时发生何事的问题。对于发现特别的有效组合，它们可能是有用的。这种组合在其他情况下可能不明显，如“噢，在我们开展电子邮件活动之后，网络标语的点击率在上升。”

17.3.3 交叉表

查看随时间变化的变量可以使用交叉表。一般说来，交叉表可以展示两个变量相对于彼此发生的频繁程度。图 17-11 显示了两个变量之间的交叉表，即渠道和信用卡支付变量。泡的大小显示在该渠道中开始采用该种支付方法的客户比例。这里的数据对应于表 17-2 所示数据。

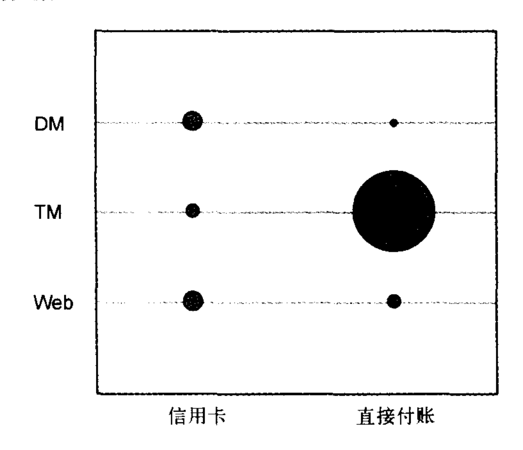


图 17-11 交叉表表示变量之间的关系

表 17-2 支付方式和渠道的交叉表

	信用卡	直接账单
DM	69 126	51 481
TM	50 105	249 208
Web	67 830	29 608

没有时间的交叉表表示静态图像而不是趋势。这是有用的，但通常来说，趋势信息更有用。

17.4 衍生变量

在本章和整本书中，已经列举许多衍生变量（derived variable）的例子。这种变量被预处理，使得数据挖掘算法将它们纳入到模型变得更容易。也许更重要的是，衍生变量使领域知识纳入数据挖掘过程成为可能。把领域信息放入数据，以便数据挖掘算法能使用它找到模式。

增加变量对任何成功的数据挖掘工程是一项核心任务，详细地查看计算衍生变量的 6 种基本方法是值得的。这 6 种方法是：

- 提取来自单个数值的特征
- 在记录内合并数值（其中包括用于捕获趋势）
- 在另外一张表中，查找辅助信息
- 选择多个列中依赖数据的主元
- 汇总交易记录
- 汇总跨越模型集的字段

以下部分讨论这些方法，给出一些衍生变量的例子，并且突出计算重点。

17.4.1 提取来自单个数值的特征

从计算上来看，因为所有需要的数据都呈现为单一数值，所以分析数值是非常简单的操作。虽然它很简单，但相当有用，如下面的例子所示：

- 从日期计算每周的某天
- 从信用卡号码提取信用卡发行者的代码
- 获得邮政编码前三个数字
- 从车辆识别码（VIN）中确定车辆制造商的代码
- 当一个字段丢失的时候，增加一个标志

这些操作常常是需要数据挖掘工具能够处理的基本操作。不幸的是，许多统计工具把重点放在数值型数据上，而不是在商业数据中时常遇到的字符串、日期和时间，因此字符串操作和日期计算可能是困难的。在这种情况下，可能需要在预处理阶段或者从数据源提取数据时增加这些变量。

17.4.2 在记录内合并数值

正如来自单一数值的特征提取，从计算角度看，在记录内合并数值也是简单的——不是使用一个变量，而是有几个变量。大多数数据挖掘工具支持增加衍生变量，合并来自几个字段的数值，特别是对于数值型字段。这对于增加比率、求和、求平均数等可能非常有用。对建模来说，这种导出数值比原始数据通常更有用，因为这些变量开始捕捉潜在客户的行为。日期字段时常被合并。取两个日期的差计算持续时间也是十分普遍和有用的例子。

通常情况下，合并字符串字段不是必要的，除非字段以某种方式相关。举例来说，将“信用卡类型”与“信用卡支付标识”结合可能是有用的，这样，就有一个字段表示支付类型。

17.4.3 查找辅助信息

查找辅助信息是比前面两种计算更复杂的过程。查找是将两张表格合并在一起（使用关系数据库的术语）的例子，按照简化的原则，一张表格是大的，另一张表格相对较小。

当查找表足够小的时候，如表 17-3 所示，它描述了信用卡号前几位数字与信用卡类型之间的映射，一个简单的公式对于查找就足够了。

表 17-3 信用卡前缀

卡类型	前缀	长度
MasterCard	51	16
MasterCard	52	16
MasterCard	53	16
MasterCard	54	16
MasterCard	55	16
Visa	4	13
Visa	4	16
American Express	34	15
American Express	37	15
Diners Club	300	14
Diners Club	301	14
Diners Club	302	14

（续）

卡 类 型	前 缀	长 度
Diners Club	303	14
Diners Club	304	14
Diners Club	305	14
Discover	6011	16
enRoute	2014	15
enRoute	2149	15
JCB	3	16
JCB	2131	15
JCB	1800	15

比较常见的情形是带有信息的次级表格或文件。举例来说，这张表格可能包含：

- 邮政编码区域的人口和中值家庭收入（由美国人口普查局 www.census.gov 提供，供美国人下载）。
- 产品代码的分层。
- 商店的零售位置类型信息。

不幸的是，对数据挖掘工具来说，通常没有编程，查找就比较困难。一些工具的确提供这种便利，如来自 Insightful 公司的 I-Miner，通常需要两个表格都要按照查找字段进行排序。图 17-12 展示了一个这样的例子。对于一个这样的字段它是令人满意的，但是当需要查找许多不同的字段时它就不方便了。大体上，在工具之外进行这些查找是比较容易的，尤其是当查找表与初始数据都来自数据库的时候。

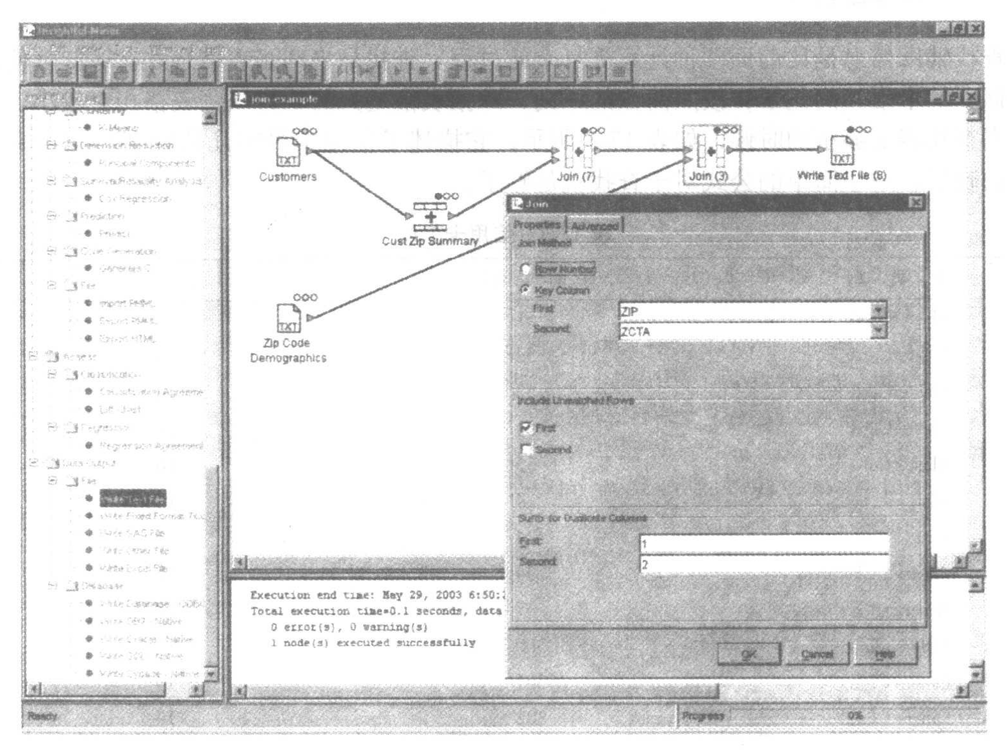


图 17-12 Insightful Miner 使用户能够从图形用户界面使用并且创建查找表

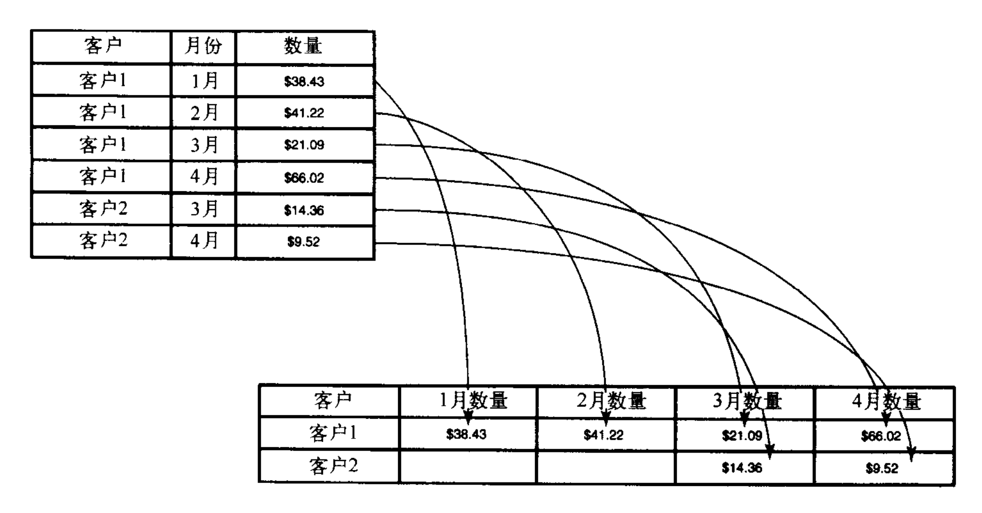
有时，查找表已经存在。而有些时侯，要根据需要创建。举例来说，一个有用的客户流失预报器是按照邮政编码统计的历史流失率。将这一点增加到客户特征标识，需要对每个邮政编码计算历史流失率，然后将结果作为查找表。

警告： 当使用数据库连接在查找表中查找数值时，总是使用左外连接，确保在这个过程中没有任何客户行丢失！在 SQL 中，一个外部连接如下：

```
SELECT c.*,l.value
FROM (customer c left outer join lookup l on c.code = l.code)
```

17.4.4 转轴正则时间序列

客户数据时常按月存储，每个月有独立的数据行。例如，由于大多数基于订阅的公司每月一次向客户发放账单，账单数据时常以这种方式存储。如果数据按照固定的、定义好的区间发生，这笔数据就是正则时间序列的例子。图 17-13 举例说明把这笔数据放入客户特征标识的过程。数据必须被转轴，以便开始以行组织的数值最后以列组织。



客户	月份	数量
客户1	1月	$38.43
客户1	2月	$41.22
客户1	3月	$21.09
客户1	4月	$66.02
客户2	3月	$14.36
客户2	4月	$9.52

客户	1月数量	2月数量	3月数量	4月数量
客户1	$38.43	$41.22	$21.09	$66.02
客户2			$14.36	$9.52

图 17-13　对于每位客户，转轴字段是取存储在一行或多行的数值，把它们置于每位客户的一行，但却在不同的列中

这通常是一个很麻烦的过程，因为数据挖掘工具和 SQL 都不能容易地进行转轴操作。数据挖掘工具常常需要编写程序进行转轴操作。为了完成这个任务，客户文件需要按照客户身份标识排序/分类，并且账单文件需要按照客户身份标识和账单日期排序/分类。然后，需要编写特定代码来计算转轴列。在 SAS 中，“proc TRANSPOSE”命令就是用于这个目的。后面“用 SQL 转轴数据”部分展示了在 SQL 中如何进行转轴操作。

大多数企业以月为基础存储客户数据，通常是按照历法月份。但有些行业显示强烈的周周期模式，因为客户在周末或做或不做事情。举例来说，网站可能在每周工作日期间是最活跃的，报纸订阅通常在星期一或星期日开始。

因为有些月份比其他月份的时间更长，所以周周期干扰月数据。考虑一个大多数活动在每周工作日进行的网站。有些月份有 20 个工作日，其他月份最多的有 23 个（不包括假日）。相连的两个月之间的差可能是 15%，这仅仅由于周工作日的数量差别。考虑到这一点，将

每月活动除以周工作日数，可以获得“每周工作日活动”。但是，当有很明显的周周期特征时，这才有意义。

用 SQL 转轴数据

对于转轴数据，SQL 没有很强的支持（虽然某些数据库可能对这种能力有非标准的扩充）。然而，当使用标准 SQL 的时候，转轴数据也是可能的。

假设数据由账单记录组成，并且每个被分配一个连续的账单号码。第一笔账单记为“1”，第二笔记为“2”，依此类推。下列 SQL 片段表示该如何转轴这笔数据：

```
SELECT customer_id,
        sum(case when bill_seq = 1 then bill_amt end) as bill_1,
        sum(case when bill_seq = 2 then bill_amt end) as bill_2,
        sum(case when bill_seq = 3 then bill_amt end) as bill_3,
        ...
FROM billing
GROUP BY customer_id
```

这个片段的一个问题是不同的客户有不同的账单周期数。然而，查询只能取固定的数字。当客户的账单周期数比查询需要更少的时候，较迟的周期用 NULL 来填充。

实际上，因为客户特征标识需要最近的账单周期数——比如说，最后的 12 或 24 个，所以这个代码片段通常不是客户特征标识所需要的。对于活跃的客户来说，这是最近的周期。然而，对于已经停止的客户，我们需要考虑他们的停止日期。下列代码片段考虑了这一点：

```
SELECT customer_id,
        sum (case when trunc(months_between(bill_date,cutoff) = 1
                then bill_amt else 0 end) as bill_1,
        sum(case when trunc(months_between(bill_date,cutoff) = 2
                then bill_amt else 0 end) as bill 2,
        ...
FROM billing b,
      (select customer_id,
              (case when status = 'ACTIVE' then sysdate
                    else stop_date end)as cutoff
       from customer) c
where b.customer_id = c.customer_id
GROUP BY customer_id
```

这个代码片段确实使用某些 SQL 的扩充来计算日期（在这个例子中，这些被表示为 Oracle 函数）。然而，大多数数据库有相似的函数。

上述代码是一个杀手查询的例子，因为它用一张大的表格（客户表格）连接一张更大的表格（客户账单表格），然后进行分组操作。幸运的是，现代数据库能很好地使用多个处理器和多个磁盘，在合理的时间内完成这个查询。

17.4.5 汇总交易记录

交易记录是非正则时间序列的例子，即记录会在任何时间点随时发生。这种记录由客户交互作用而产生，现实案例有：

- 自动柜员机交易
- 电话呼叫
- 网站访问
- 零售

当采用非正则时间序列开展工作的时候，就遇到几个挑战。首先，交易量非常大。在如此海量的数据上工作需要复杂的工具和强有力的计算机。其次，没有标准的方法用于这项工作。正则时间序列数据有自然的转轴方法。而对于非正则时间序列而言，决定如何最好地汇总数据是必需的。

一种方法是，把非正则时间序列转变成正则时间序列，然后转轴序列。举例来说，计算每个月呼叫的数量，或者每个月在自动柜员机上提款的数量，然后按月转轴总数。当处理交易时，这些计算可能更复杂，例如长度超过 10 分钟的呼叫，或者低于 50 美元的提款。这些特殊的汇总可能是相当有用的。描述客户行为的更复杂例子将在下一节之后提供。

另一种方法是定义一组数据变换，在收集交易数据时运行。这是电信行业所使用的一种方法，其中数据量是巨大的。某些变量可能是像使用的分钟数一样简单，而有些可能像呼叫号码是企业号码还是居民号码的评分一样复杂。这种方式使代码计算非常困难，并且这种计算是很难改变的。尽管这种变量可能有用，但比较有弹性的环境对于汇总交易数据从策略上来说更有用。

17.4.6 汇总跨越模型集的字段

对于衍生变量，最后的方法是汇总客户特征标识本身字段的值。有几个这种字段的例子：

- 将数值分到同等大小的箱子中，需要计算箱子的拆分点。
- 标准化数值（减去均值，并且除以标准差），需要计算字段的均值和标准差，然后再进行计算。
- 排列数值（最小的数值为 1，第二小的数值为 2，依此类推）需要排序所有的数值以获得分级。

虽然这些操作很复杂，但是它们都直接在模型集上运行。数据挖掘工具为这些操作提供支持，尤其是对三者中最重要的分箱数值型数值。

可能非常有用的一类分箱不容易得到，那就是基于频率对代码进行分箱。例如，在模型集中保存至少 1000 个实例的所有代码，把所有其余的代码放在单独的“其他”类中，这将是有用的。这对于处理离群值是有用的，如在电话数据中那些旧的、不流行的移动电话，虽然少数客户仍使用它们。一种处理方法是，标识要保存的移动电话，即增加新的字段“要分析的移动电话”保存这些移动电话，并把其余的放进一个“其他”类中。更自动的方法是创建查找表来映射这些移动电话。然而，也许更好的方法是用诸如移动电话发布日期、权重和使用特征等信息替换移动电话 ID，这些信息可能在查找表中已经可用。

17.5 基于行为变量的例子

衍生变量的真正力量来自它沿着已知维度汇总客户行为的能力。本部分构建已经展示的想法，并且给出三个有用的基于行为变量（behavior-based variable）的例子。

17.5.1 购买频率

从前，目录编辑设计了一种巧妙的方法，使用三个维度刻画客户行为，即崭新度（R）、频率（F）和消费金额（M）。基于这三个变量的 RFM 至少自 20 世纪 70 年代以来就已经被使用。客户行为的这三种描述中，崭新度通常是最具预言性的，但频率是最值得关注的。崭新度只是意味着客户自购买以来的时间长度。比较传统的是，消费金额是购买的总量（虽然我们已经发现，由于总数与频率高度相关，平均购买量更有用）。

在传统的 RFM 分析中，频率只是购买的次数。然而，简单的计数不能很好地刻画客户行为。有一些其他确定频率的方法，并且这些方法可以应用于与目录购买不相关的其他领域，包括抱怨频率、打国际长话的频率，等等。重要的是，客户可能在不规则的时间区间完成行为，我们之所以要刻画这种行为模式，是因为它提供关于客户的潜在有用的信息。

计算频率的一种方法是获得历史数据给出的时间长度，然后除以客户购买的次数。因此，如果目录数据追回到 6 年前，并且客户只进行了一次购买，那么频率就是每 6 年一次。

这个方法尽管简单，却丢失了重要的一点。考虑下面两位客户：

- 约翰在 6 年以前有一次购买，并且从此以后收到每个目录。
- 玛丽刚刚在上个月进行了一次购买，那时她第一次收到目录。

认为这两位客户有相同的频率有道理吗？答案是不。很明显约翰的频率是每 6 年不超过一次，而玛丽仅仅在上个月才有机会进行购买，因此，她的频率的更精确描述应该是每月一次。关于频率首要的一点是，应该从客户有机会购买的那一点进行测量。

还存在另一个问题。关于约翰和玛丽，我们真正知道的是他们的频率分别是不超过每 6 年一次和每月一次。从历史角度看，一次观察不足以得出真正的频率。这实际上是一个时间与事件问题，就像在第 12 章中讨论的一样。

我们这里的目标是用衍生变量刻画频率，而不是预测下一事件（使用生存分析是最好的途径）。为了达到这一目的，假设有两个或更多的事件，事件之间的平均时间是总的时间间隔除以事件数减 1，如图 17-14 所示。它提供了在事件发生期间，事件之间的平均时间。

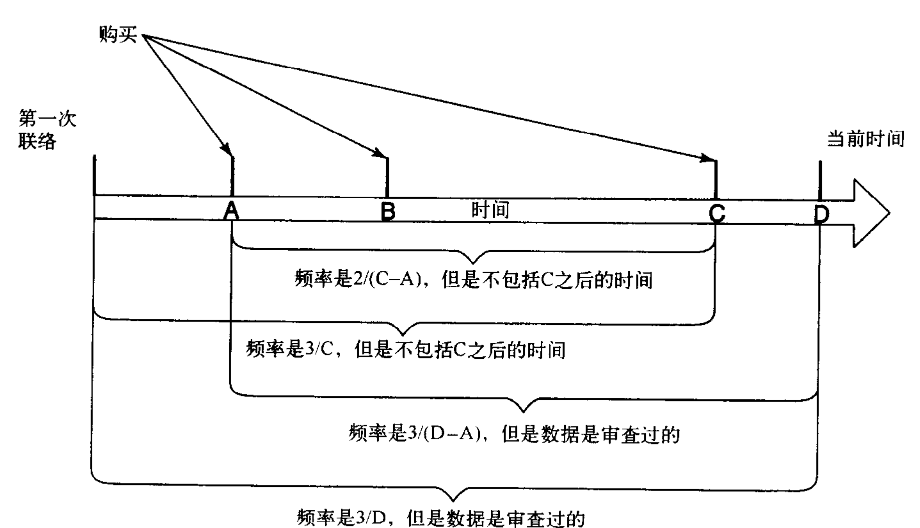


图 17-14 没有完美的方法来估算频率，但这四种方法都是合理的

对频率问题没有完美的解决办法，因为客户事件无规律地发生，而且也无法预知在未来数据被审查时会发生什么事情。取第一个事件到最近事件的时间间隔引发一个问题，即所有事件都在很久以前发生的客户可能有很高的频率。可以选择的办法是取第一个事件发生以来的时间，本质上假定当前是一个事件。这没有问题，因为下一个事件是未知的，而且当处理审查数据时必须小心。实际上，取第一件事件以来总时间间隔（或客户活跃的时间间隔）除以可能发生的事件数是最好的解决方法。

17.5.2 衰减使用

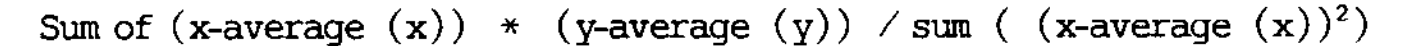
在电信行业中，流失的重要预报器是衰减使用（declining usage），即随着时间的过去，使用服务越来越少的客户比起其他客户更有可能离开。有衰减使用的客户可能有许多变量指示这一点：

- 账单度量，例如最近的花费数量相当小。
- 使用量度量，例如最近的使用量相当小，或者每月总是最小量。
- 最近没有使用可选服务。
- 最近度量和旧的度量的比率小于 1，并且时常远远小于 1，表示最近使用比历史使用小。

对同样的潜在行为存在众多不同的度量，暗示了一种情形，即衍生变量以单个变量的形式可能有益捕捉行为。目标是尽可能将很多信息纳入“衰减使用”的指示器。

· **提示：** *当许多不同的变量都指示单一的客户行为时，合并这类信息的衍生变量可能对数据挖掘更有益。*

幸运的是，数学提供了优美的解决办法，它采用最佳拟合线的形式，如图 17-15 所示。拟合的好坏程度用 R^2 统计量描述，变化范围从 0 到 1，数值靠近 0 代表差的拟合，靠近 1 代表好的拟合。线的倾斜度说明，某一变量随时间的平均增加率或减少率。在统计学上，该倾斜度称为 Beta 函数，并且按照下列公式进行计算：



Sum of (x-average (x)) * (y-average (y)) / sum ((x-average (x))²)

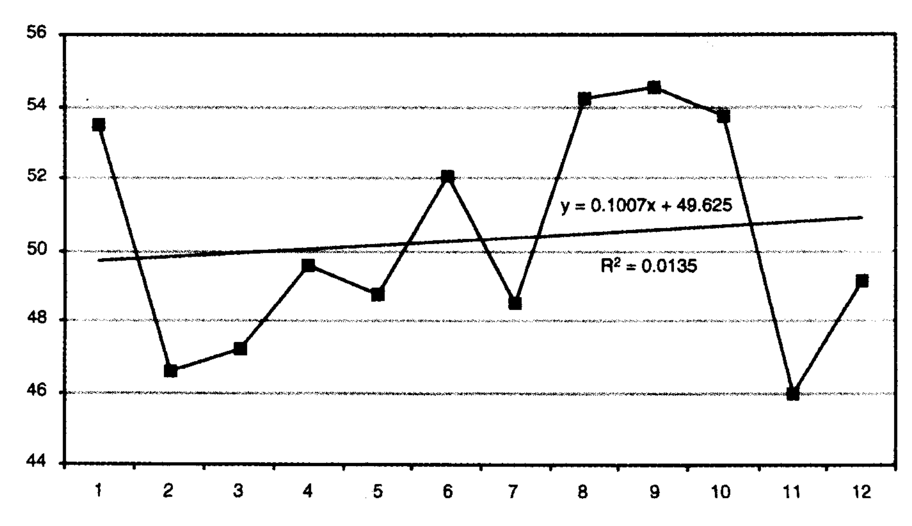


图 17-15　最佳拟合线段的斜率提供一个很好的按时间变化的度量

为给出如何使用这个公式的例子，考虑图 17-15 所对应的下列客户数据。表 17-4 中的数据是对一个典型客户的计算。

表 17-4 计算时间序列斜率的例子

月 (X 值)	X-AVG (X)	(X-AVG (X))^2	Y (来自客户 A)	Y-AVG (Y)	(X-AVG (X))* (Y-AVG (Y))
1	−5.5	30.25	53.47	3.19	−17.56
2	−4.5	20.25	46.61	−3.67	16.52
3	−3.5	12.25	47.18	−3.10	10.84
4	−2.5	6.25	49.54	−0.74	1.85
5	−1.5	2.25	48.71	−1.57	2.35
6	−0.5	0.25	52.04	1.76	−0.88
7	0.5	0.25	48.45	−1.83	−0.91
8	1.5	2.25	54.16	3.88	5.83
9	2.5	6.25	54.47	4.19	10.47
10	3.5	12.25	53.69	3.42	11.95
11	4.5	20.25	45.93	−4.35	−19.59
12	5.5	30.25	49.10	−1.18	−6.51
总计		143			14.36
斜率					0.1004

该例展示了计算斜率的一个典型应用，即发现上一年的使用情况或者账单模式的上升情况。表格式用最适合电子数据表的格式显示计算情况。然而，许多数据挖掘工具提供一个函数，直接从一行中的一组变量计算 Beta 数值。当这种函数不可用的时候，可以使用更基本的算术函数来表达。

虽然对于这种计算，每月数据时常是最方便的，但是要记住不同的月份天数不同。这个问题对于有强烈的周周期特性的商业特别重要。举例来说，有些月份有 5 个完整的周末，而其他的只有 4 个。不同的月份有 20～23 天的工作日（不包括节假日）。这些差异占月份之间差异的 25%。当处理有这类周期的数据时，一个好主意是计算“每周末的平均值”或“每个工作日的平均值”，以便考察所选度量如何随时间变化。

提示：当利用有周周期，但必须按月报告的数据时，可以考虑诸如“周末每天的平均值”或“工作日每天的平均值”之类的变量，这样月份之间的比较就更有意义。

17.5.3 旋转者、交易商和便利用户：定义客户行为

时常，经商者能够基于客户随时间变化的行为刻画不同的客户群体。然而，将非正式的交易描述翻译成对数据挖掘有用的形式具有挑战性。面对这种挑战，最好的对策是确定与商业理解相匹配的客户行为度量。

本例是关于在主要零售银行的信用卡群体，已经发现可盈利客户有三种风格：

- *旋转者*是在信用卡上维持大宗收支差额的客户。因为他们每个月为巨大的收支差额支付利息，所以是高利润客户。
- *交易商*是每个月有大宗收支差额，但会全部付清的客户。这些客户不支付利息，但是对每笔交易收取的交易费是重要的税收来源。交易费的一个组成部分以交易量的百分

比为基础。

- 便利用户是定期借用大量费用的客户，例如，为了度假或大宗购买，然后在几个月内付清。虽然不像旋转者那样盈利丰厚，但他们的风险较低，同时要支付大量的利息。

市场营销组相信，这三类客户受不同需求所驱使。因此，了解未来客户行为，就可能允许未来的营销活动将最适当的信息发送给每个客户片段。群体要预测未来 6 个月的客户行为。

本例的关键部分不是预测，而是片段的定义。训练集需要把客户已经分为三个组的例子。获得这个分类（classification）被证明是一个挑战。

1. 数据

这个工程可用的数据由 18 个月的账单数据组成，包括：

- 信用额度
- 利率
- 每个月新收取的费用
- 最小支付量
- 已付的数量
- 每个月的总余款
- 每个月已付的利息和相关的费用

这是信用卡的典型规则。当客户已经还清余款的时候，对新的收费无需支付利息（1 个月的时间）。然而，当有很大的余额时，对余额和新的收费都要支付利息。这笔数据对了解客户有什么启发？

2. 根据估计收益进行分段

估计收益是理解客户价值的好方法（本质上，该数值对客户行为不提供很多的深入了解，因此对于宣传并不是很有用）。单独以客户价值和其收入为基础，假设所有客户的花费是相同的。尽管这不是事实，却是一个有用的近似值，因为一个完美的收益模型是相当复杂的，而且很难开发，已经超出本例的范围。

表 17-5 给出了 6 位客户 1 个月的账单。最后一列是估计收益，有两个组成部分。第一个是支付的利息量，第二个是新交易的交易费用，在本例中这个估计值是新交易量的 1%。

表 17-5 六位信用卡客户及其一个月的数据

	信用额度	利率	新的收费	初始余额	最小支付	付款量	利息	交易收入	估计收益
顾客 1	$500	14.9%	$50	$400	$15	$15	$4.97	$0.50	$5.47
顾客 2	$5 000	4.9%	$0	$4 500	$135	$135	$18.38	$0.00	$18.38
顾客 3	$6 000	11.9%	$100	$3 300	$99	$1 000	$32.73	$1.00	$33.73
顾客 4	$10 000	14.9%	$2 500	$0	$0	$75	$0.00	$25.00	$25.00
顾客 5	$8 000	12.9%	$6 500	$0	$0	$6 500	$0.00	$65.00	$65.00
顾客 6	$5 000	17.9%	$0	$4 500	$135	$135	$67.13	$0.00	$67.13

估计收益是用单个数值比较不同客户的好方法。这个表格清楚地说明，很少使用信用卡（客户 1）的人，估计收益也很少。另一方面，缴纳很多收费或支付利息的人产生较大的收益。

然而，估计收益不能区分不同类型的客户。事实上，交易商（客户 5）有非常高的收益，没有新费用（客户 6）的旋转者也是如此。该例显示估计收益与客户行为的关系很小。频繁使用信用卡的用户和很少使用的用户都产生很多收益。这是我们所期望的，因为有不同类型的盈利客户。

真实世界比这个简单例子更复杂。每位客户都有破产的风险；那样，很突出的余额一定被勾销。不同类型的卡有不同的规则。举例来说，许多合作的卡要给合作机构支付交易费。并且，服务不同的客户花费也不同，取决于客户是否使用客户服务、投诉收费、在线支付，等等。

简而言之，估计收益是了解哪些客户有价值的好方式。但是，对客户行为不提供更多深入了解。

3. 根据潜能分片

除了实际收益之外，每位客户都有潜在收益（potential revenue）。这是客户每个月可能会产生的最大收益量。最大收益容易计算。简单假设整个信用卡用于新的收费（因此产生交易税）或者结转（利息税收）。这些中较大的就是潜在收益。

表 17-6 比较了在一个月期间，同样 6 位客户的潜在收益和实际收益。这张表格展示了一些有用的特征。某些不盈利的客户已经达到潜能的饱和状态。不增加信用限度或利率，是不可能增加他们的价值的。

表 17-6 六位信用卡客户的潜力

	信用额度	利率	利息	交易	潜在收益	实际收益	潜力
顾客 1	$500	14.9%	$6.21	$5.00	$6.21	$5.47	88%
顾客 2	$5 000	4.9%	$20.42	$50.00	$50.00	$18.38	37%
顾客 3	$6 000	11.9%	$59.50	$60.00	$60.00	$33.73	56%
顾客 4	$10 000	14.9%	$124.17	$100.00	$124.17	$25.00	20%
顾客 5	$8 000	12.9%	$86.00	$80.00	$86.00	$65.00	76%
顾客 6	$5 000	17.9%	$74.58	$50.00	$74.58	$67.13	90%

比较实际收益和潜在收益有另一个方面的问题：使数据归一化。没有归一化，较富有的客户似乎有最大的潜能，尽管这个潜能没有被完全利用。因此，具有 $10 000 信用额度的客户远远不能达到他或她的潜能。事实上，客户 1 有最小的信用额度，最可能达到他或她的潜在价值。这种价值定义排除富有因素，可能未必适合特定的目的。

4. 与理想情况比较，获知客户行为

因为估计收益和潜能不能区分客户行为的类型，让我们回过头来更详细地看看定义。首先，在数据内部，是什么说明谁是旋转者？下面是一些旋转者的定义：

- 每个月支付利息的人
- 每个月支付超过特定量利息的人（譬如说，超过 $10）
- 几乎每个月支付超过特定量利息的人（譬如说，在 80%的月份中超过 $10）

所有这些都有一个特别的性质（并且营销群体在历史上已经给出类似的定义）。那些支付很少利息，但是每个月都支付利息的人怎么样呢？为什么是 $10？为什么是 80%的月份？这些定义都是任意的，通常是一个人在特定时间对定义的最好推测的结果。

从客户角度看，旋转者是什么？是每个月只有最小付款量的人。到现在为止，一直都还不错。对于比较客户，这个定义有点不准确，因为最小付款按月份和客户的变化而改变。

图 17-16 显示三位客户发生的实际付款和最小付款，他们的信用额度全都是 $2 000。旋转者每个月的付款非常接近最小付款。交易商付款比较接近信用额度，但是这些每月费用变化幅度非常大，这取决于每个月发生的缴费量。便利用户差不多是在二者之间。从性质上看，曲线的形状对客户的行为提供了深入了解。

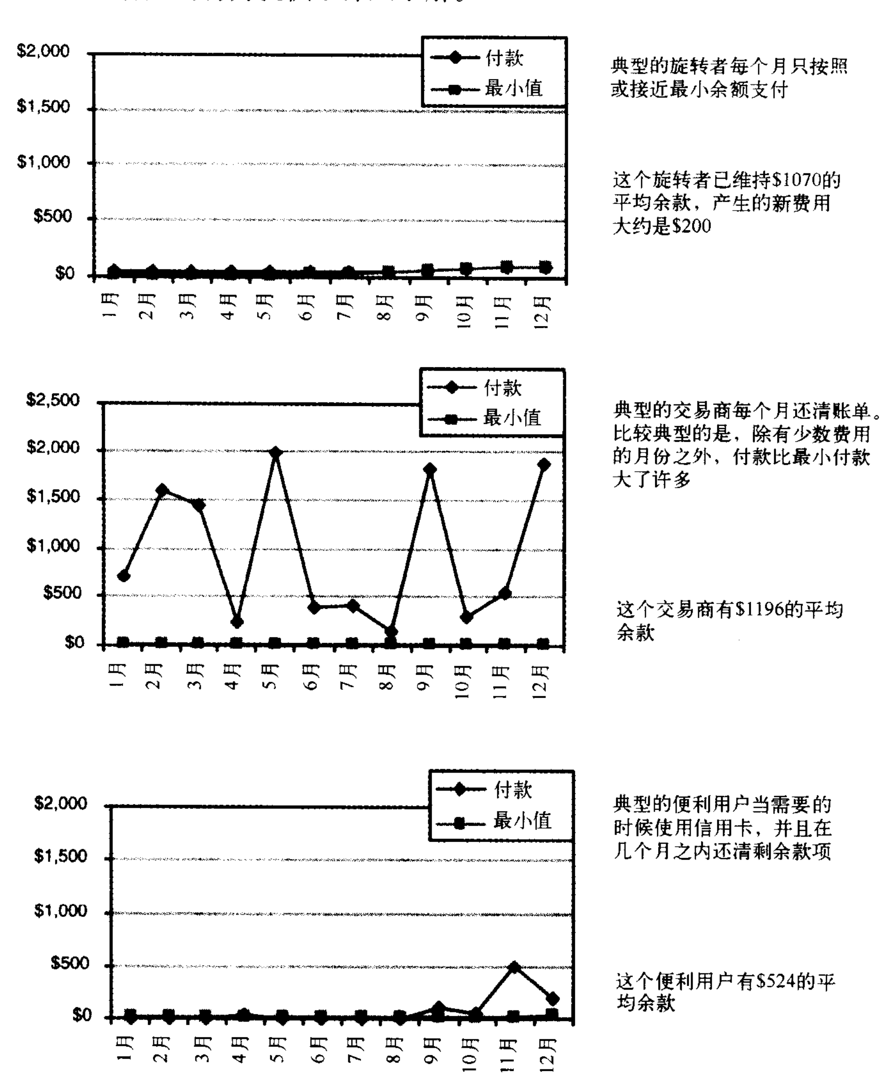


图 17-16 这三个图展示了拥有 $2000 信用额的三位信用卡客户实际的和最小付款

当分类数百万客户行为时，手工描绘曲线是效率很差的一种方法。曲线是一种含糊的定性概念，真正需要的是一个得分。产生得分的方法是考察“最小付款”曲线和实际“付款”曲线之间的区域。就我们的目的来说，该区域是付款和最小量之间差异的总和。对于旋转者来说，这个总数是 $112；对于便利用户来说，是 $559.10；对于交易商，却是巨大的 $13 178.90。

这个得分有直观的意义。得分越低，该客户看起来越像旋转者。然而，得分不能比较两个拥有不同信用额度的持卡人。考虑一种极端的情况。如果持卡人有 \$100 的信用额度，并且是一个理想的交易商，那么，得分不会超过 \$1200。而拥有信用额度为 \$2000 的非理想旋转者仍然有非常大的得分。

解决的办法是，将每个月的差额除以总的信用额度，使数值归一化。现在，三者的得分分别是 0.0047、0.023 和 0.55。当归一化的得分接近 0 的时候，持卡人接近理想旋转者。当得分接近 1 的时候，持卡人接近理想的交易商。在二者之间的数字代表便利用户。这为每位客户提供了旋转者－交易商得分，便利用户位于中间。

客户行为的这个得分有一些有用的性质。从来不使用信用卡的人会有最小付款量 0，实际付款量也是 0。这些人看起来像是旋转者。那可能不是件好事。解决这个问题的一种方法是包括带有行为得分的估计收益潜能，实际上，使用两个数字描述行为。

这个得分的另一个问题是，随着信用额度的增加，客户越来越显得像旋转者，除非客户缴费更多。为了避免这个问题，比率可以改为每月余款除以信用额度。当无亏欠和无支付的时候，所有的数值都是 0。

图 17-17 显示了这个问题的一种变形。得分使用支付量对最小付款量的比率。它有一些很好的特征。理想旋转者的得分是 1，因为他们的支付和最小付款量相等。不使用信用卡的人得分为 0。交易商和便利用户两者的得分都超过 1，但是很难区分他们。

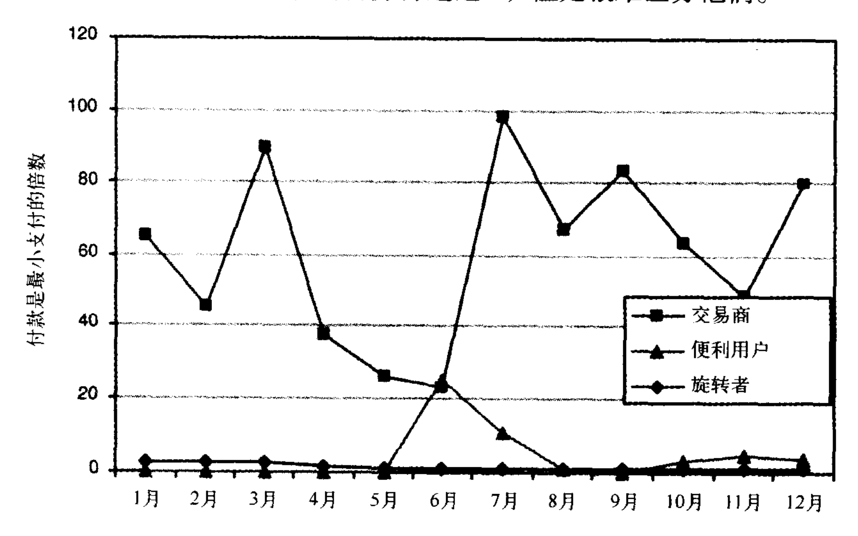


图 17-17 比较支付量与最小付款的倍数显示，对于交易商、旋转者和便利用户有截然不同的曲线

这一节已经展示了测量客户行为的几种不同的方法。所有这些方法都是基于与客户相关的重要变量和基于几个月的度量。不同的测量对识别行为的不同方面更有价值。

5. 理想的便利用户

前一节的度量重点关注客户行为的极端情况，如有代表性的旋转者和交易商。便利用户仅仅被假定差不多在中间的某个位置。是否有方法开发一个最适用于理想便利用户的得分？

首先，让我们定义理想的便利用户。某人一年两次使用信用卡达到信用额度，然后在 4 个月期间还清透支款。在那年其余的 10 个月期间，即使有，也是很少的额外费用。表 17-7 显示了两个便利用户每月的余款与信用额度的比率。

表 17-7　按照信用额度百分比表示两个便利用户的月度余额

	1月	2月	3月	4月	5月	6月	7月	8月	9月	11月	12月
Convl	80%	60%	40%	20%	0%	0%	0%	60%	30%	15%	70%
Conv2	0%	0%	83%	50%	17%	0%	67%	50%	17%	0%	0%

这张表格也阐明在便利用户定义中的主要挑战之一。描述他们行为的数值在任何给定的月份中彼此没有关系。这是不协调的。事实上，一方面便利用户之间有基本的差异，交易商和旋转者是另一方面。知道某个人是交易商，可以准确地描述他们在任何给定月份的行为，即他们会还清剩下的余额。知道某个人是便利用户没有多大的帮助，在任何给定的月份中，他们可能没有任何支付，或还清所有债务或部分还款。

这意味着不可能开发度量以识别便利用户吗？一点也不。解决的方法是，按照余额比率将 12 个月的数据进行排序，使用排序数据创建便利用户的度量。

图 17-18 举例说明这一过程。它展示了两个便利用户和理想便利用户的线轮廓。这里，数据已经排序，最大的数值首先出现。对于第一个便利用户，1 个月指的是一月。对于第二个，它指的是三月。

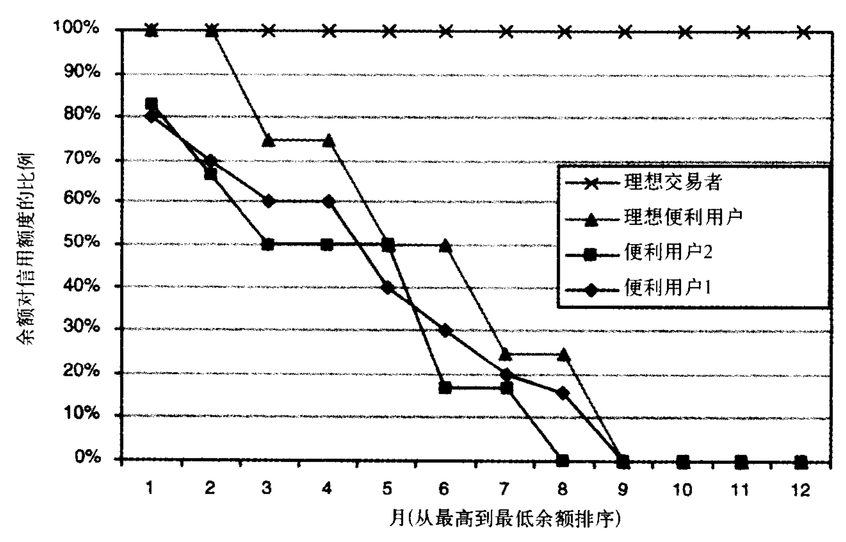


图 17-18　通过对月份余额比率排序，比较两个便利用户与理想便利用户

现在，使用同样的观念，取理想线条和实际线条之间的区域产生得分，测量便利用户接近理想状态的程度。值得注意的是，对所有月份旋转者在最大值附近会有很突出的余额。他们会有很高的得分，表明他们离理想便利用户很远。对于便利用户，得分很小。

这一案例研究已经显示划分客户的几种不同方式，都是利用衍生变量描述客户行为。通常，描述特定行为，然后创建一个得分以测量客户行为与理想情况相差多远，这种做法是可行的。

17.6　数据的黑暗面

处理数据是数据挖掘过程的关键部分。数据意味着什么？有许多方法来回答这个问题——通过书面文件，在数据库模式中，在文件布局中，通过元数据系统，而且相当重要的

一种途径是，通过了解实际发生事情的数据库系统管理员和系统分析员。无论文档编制如何好，真正的内情在数据之中。

有一种误解认为数据挖掘需要完美数据。在商业分析界，完美绝对是足够好的敌人。原因之一是，探索数据和构建模型能够突出用其他方式不知道的数据问题。利用可用数据开始挖掘过程，未必产生最好的模型，但是它确实启动了一个能随时间逐渐完善的过程。原因之二是，等待完美数据时常会延误工程，从而无法完成任何任务。

本节介绍了一些重要的问题，这些问题使处理数据成为一个痛苦的过程。

17.6.1 缺失值

缺失值指的是应该有但却没有的数据。在许多情况下，缺失值在数据源中用 NULL 来表示，很容易识别。然而，要小心的是：NULL 有时是可接受的数值。在这种情况下，我们说数值为空而不是缺失，尽管在源数据中两者看起来相同。举例来说，账户的停止代码可能是 NULL，表明账户仍然是活跃的。这个信息表明数据是否被审查过，它对生存分析至关重要。

有时 NULL 是可接受的数值，此时利用重叠数据描述客户和潜在客户的人口统计学特征和其他特征。在这种情况下，NULL 时常有下列两种意思之一：

- 没有充足的证据表明该字段对个体是否为真。举例来说，没有订阅高尔夫球杂志意味着此人不是打高尔夫球的人，但是不能证明它。
- 在重叠数据中，此人没有与之相匹配的记录。

提示：当处理重叠数据时，用另外的数值代替 NULL 是有用的，一个代表记录不匹配，另一个代表数值是未知的。

区分这些情形是有用的。一种方法是分开记录不匹配的数据，创建两个不同的模型集。另一种方法是用另外的数值代替 NULL，指出匹配失败是在记录层次还是在字段层次。

因为客户特征标识使用如此多的聚集数据，对于各种不同的特征时常包含“0”。因此，对于算法来说，在客户特征标识中缺失数据（missing data）不是最重要的问题。然而，这可能太离谱。考虑有 12 个月账单数据的客户特征标识。在过去 12 个月中开始的客户已经失去较早几个月的数据。在这种情况下，用某个任意值代替缺失数据不是一个好主意。最好的办法是把模型集拆分为两部分，即一个模型集包含有 12 个月保有期的客户，另一个包含最近的客户。

当缺失数据是问题的时候，重要的是找出原因。举例来说，我们曾经遇到一个数据库缺失客户开始日期的数据。经过进一步的调查，发现这是那些在 1999 年 3 月之前开始并结束关系的所有客户。这个数据来源后来关注这个日期之后开始的客户，或者在此日期仍是活跃的客户。在另一种情形中，交易表格丢失了某个日期之前的特定交易类型。在创建数据仓库期间，不同交易在不同时间完成。只有每次都小心地观察交叉表中的交易类型，才会弄清楚某一种类型比其余类型更晚完成。

在另一种情形中，数据仓库中的缺失数据就那样缺失了，因为数据仓库不能适当地载入它。当原因清楚的时候，应该修复数据库，因为误导数据比没有任何数据更糟。

处理缺失数据的一种方法是试着填写数值，例如利用平均数值或最常见的数值。任何一个替换值会改变变量的分布，并且可能导致产生拙劣的模型。这个方法的一种比较聪明的变

体是，使用像回归或神经网络之类的技术，试着计算基于其他字段的数值。除非绝对需要，否则我们不赞成这种方法，因为字段不再意味着它应该意味的事情。

警告： 处理缺失数值最糟糕的方法之一是用某些“特别的”数值替换它们，例如9999或-1，由于它的不合理性，会很醒目。数据挖掘算法将会恰当地使用这些好像为真的数值，导致不正确的结果。

通常，数据缺失是由于系统原因，就像在前面提及的新客户情形。较好的方法是将模型集分拆为几个部分，从一个数据集中消去缺失字段。虽然一个数据集有多个字段，但都不再有缺失值。

了解数据在未来是否缺失也很重要。有时，正确的方法是在有完整数据的记录（希望这些记录完全代表所有的记录）上建立模型，并且让人修复数据源，消除未来隐患。

17.6.2 脏数据

脏数据指的是包含可能看起来正确但实际不正确的数值的字段。通常可以标识这些数据，因为这类数值是离群值。例如，从前一家公司认为，呼叫中心的接线员收集客户的出生日期是非常重要的，所以他们将荧屏上的相应输入栏设为强制性的。当他们观察数据的时候，惊讶地发现超过5%的客户生于1911年，并且不仅仅在1911年，而且在11月11日。事实上，不是所有客户都在这一日期出生，呼叫中心的接线员很快获知打6个“1”是填充该字段（日、月和年，每个填两个字符）的最快方法。结果是：许多客户拥有恰好相同的生日。

收集精确数据的企图时常与管理企业的工作相冲突。许多商店对有会员卡的客户打折。当客户没有会员卡的时候，会发生什么呢？商业规则或许说“没有折扣”。可能真正发生的是，商店职员可以敲入默认数字，所以客户仍然能取得资格。这种友好的举止导致某个会员号码似乎有格外高的交易量。

一家公司发现几位客户在新泽西州伊丽莎白市，其邮政编码为07209。不幸的是，当按照邮政编码和附加邮政编码信息分析数据的时候，发现这个邮政编码根本不存在。因为邮局时常能解决发送写错地址的邮件，所以早期没有发现这个错误。此类错误能通过使用软件或外部服务局标准化地址数据来修复。

实际上，看起来像脏数据的数据可能实际上提供了对商业的深入了解。例如，电话号码应该只能由数字组成。一家地方电话公司的账单系统把数字存储成字符串（实际上，这是相当普遍的）。令人惊讶的是，有几百个“电话号码”包括了字母字符。在被问及此事之后几个星期（!），系统组确定，本质上这些是电话卡号码，不能附到电话线路，只能为第三方账单服务使用。

另外一家公司使用媒体代码决定如何获得客户。因此，以“W”开始的媒体代码表示客户来自网络，“D”表示对直接邮寄的响应等。在代码中，另外的字符用来区分特别的标语广告和特别的电子邮件活动。当观察数据的时候，人们惊讶地发现网络客户都开始于80年代。不，这些不是早期客户。事实上媒体代码的编码方案在1997年10月创建。较早的代码本质上是乱语。解决方法是创建新的渠道用于分析，即“1998年之前”的渠道。

警告： 大多数有害的数据问题是你不知道的。由于这个原因，数据挖掘不能在真空中进行；来自商业人士和数据分析师的输入对于成功是至关重要的。

所有的这些案例情形是能识别脏数据的例子。然而，数据挖掘的最大问题是那些未知的情况。有时，数据问题隐藏于系统干预的背后。特别是，某些数据仓库建立者憎恶丢失数据，因此，在清理数据时，他们可能归于一些数值。举例来说，在 1998 年，一家公司有一半以上的忠实客户注册了公司的忠诚卡计划。计划大概已经执行很久了，但是数据是在 1998 年装入数据仓库中。猜猜发生了什么事？对于首次载入的客户，数据仓库建立者只是加入了当前的日期，而不是客户实际注册的日期。

数据挖掘的目的是找到数据中的模式，最好是令人感兴趣的、可操作的模式。最明显的模式以公司如何运转为基础。通常情况下，目标是获得对客户的了解，而不是对商业如何运行的了解。为了做到这一点，有必要了解创造数据的时候正在发生的事情。

17.6.3 不一致数值

从前，计算机是昂贵的，因此公司没有很多的计算机。那个时代已过去很久，现在公司有许多系统以满足不同的需求。事实上，大多数公司有数十或数以百计的系统，一些立足于操作层面，一些立足于决策支持层面。在这样的一个世界中，不可避免的，不同系统的数据不总是一致的。

系统不一致的一个原因是，它们指的是不同的事情。考虑移动电话服务的开始日期。订单－登陆系统可能考虑客户签购服务的日期，操作系统可能考虑服务生效的日期，账单系统可能考虑第一份账单生效的日期。下游的决策支持系统可能还有另一种定义。所有这些日期应该彼此相靠近。然而，总有例外。最佳解决办法是包括全部的日期，因为它们能使业务更清楚。举例来说，在客户注册服务的时间和服务实际生效的时间之间，何时有长时间的延迟？这与流失有关吗？比较普通的解决办法是选择其中的一个日期，并称之为开始日期。

另一个原因与系统开发者的正确意图有关。举例来说，决策支持系统可能保存客户的当前快照，包括客户为什么停止的代码。一个代码数值可能表明某些客户因为非支付原因而停止；而某些代码数值可能表示其他理由——转向竞争对手、不喜欢服务，等等。然而，对已经自动停止的客户不支付最后账单并不罕见。在这个数据源中，实际的停止代码被简单地覆盖。客户停止的时间愈长，当公司确定一笔余款应该被归还的时候，初始停止理由随后被覆盖的机会越大。这里的问题是一个字段被用于两件不同的事情——停止理由和非支付信息。这是拙劣的数据建模反过来刺痛分析者的例子。

使用数据仓库带来的一个问题是如何区分初始载入和后来逐渐增加的数据。通常，初始载入没有丰富的信息，因此按时间追溯回去时存在空白。举例来说，初始日期可能是正确的，但是没有任何那个日期的产品或账单计划。数据的每个来源有它的特质；最好的建议是开始了解数据并提许多问题。

17.7 计算问题

创造有用的客户特征标识需要相当可观的计算能力。幸运的是，计算机担当此任。较重要的问题是使用哪一个系统。有几种进行转换工作的可能：

- 源系统，典型的是在某类数据库中（操作性或者决策支持类型）
- 数据提取工具（用于填充数据仓库和数据集市）
- 专用代码（例如 SAS、SPSS、S-Plus、Perl）

- 数据挖掘工具

其中每一个都有自己的优缺点。

17.7.1 源系统

源系统通常是关系数据库或大型机系统。通常，这些系统是高度受限的，因为它们有许多用户。这种源系统不是完成数据转换的可行平台。相反地，数据来源于这些系统（通常作为平面文件），但在其他地方进行处理。

在其他情况下，数据库可能对特别的查询有用。由于关系数据库的能力，这种查询对产生客户特征标识是有用的。特别是，数据库可能：

- 从个别字段中提取特征，即使这些字段是日期型和字符型
- 使用算术运算合并多个字段
- 在参照表中查找数值
- 汇总交易数据

关系数据库不是特别擅长转轴字段，尽管如本章前面所示，它可用于这个目的。

就缺点而言，在 SQL 中表示转换可能很麻烦，至少需要具备相当的 SQL 专用技术。查询可能扩展为数百行，包含子查询、连接以及聚合运算。这种查询尤其不易读懂，除了构造它们的人。这些查询也是杀手查询，尽管数据库正逐渐变得有力并能处理它们。从正面看，数据库确实可以充分利用并行硬件，是转换数据很有利的条件。

17.7.2 提取工具

提取工具（时常称为提取－转换－载入 ETL 工具）通常用于装载数据仓库和数据集市。在多数公司中，商业用户不能随时访问这些工具，并且它们的大部分功能能够在其他工具中发现。提取工具通常是昂贵的，因为它们是专为大型数据仓库工程而设计。

在 *Mastering Data Mining*（Wiley，1999）一书中，我们使用 Ab Initio 公司的一组工具讨论了一个案例研究。这家公司专门研究并行数据转换软件。该案例研究显示了这种软件处理大量数据的能力，也表明在这类软件可用的环境中要考虑的事情。

17.7.3 专用代码

代码是百试百灵的完成数据转换的方法。工具的选择真正以程序员熟悉的和可用的工具为基础。对于客户特征标识需要的转换，主要的统计工具都有足够的功能。

使用专用代码的一个缺点是它把额外的层加入数据转换过程。数据仍然必须从源系统（一个可能的错误源）提取，然后通过代码（另一个错误源）传递。撰写具有很好文档说明的、能重用的代码是一个不错的主意。

17.7.4 数据挖掘工具

逐渐地，数据挖掘工具有能力利用现有工具转换数据。虽然对非数值数据类型的支持因工具而不同，但大多数工具有能力从字段中提取特征，并且将多个字段结合在一起。某些工具也支持客户特征标识汇总，例如分箱变量（其中，首先通过观察整个数据集来决定分箱断点）和标准化。

然而，数据挖掘工具通常在查找数值和聚集数据方面很弱。由于这个原因，客户特征标识几乎总是在其他地方产生，然后载入工具。来自主要厂商的工具允许程序代码嵌入工具中，并且使用SQL访问数据库。使用这些特征是一个好主意，因为这些特征减少了转换数据时需要追踪的事情数量。

17.8 小结

数据为驱动数据挖掘提供动力。数据准备（data preparation）的目标是提供清洁的燃料，以便分析引擎尽可能高效地工作。对于大多数算法而言，最佳输入使用客户特征标识的形式，即一个单独的数据行带有描述客户不同方面的字段。这些字段多数是输入栏，有几个是预言性模型的目标。

不幸的是，客户特征标识与在可用的系统中发现数据的方式不同——一个很好的理由是客户特征标识随时间而改变。事实上，凭借构成有用信息的数据和主意的改变，它们经常被创建和重建。

源字段有几个不同的类型，例如数值、字符串和日期。然而，最有用的数值通常是那些附加的数值。创建衍生数值（derived value）可能像合并两个字段那么简单，或者，可能需要非常复杂的在大量数据上的计算。当试图根据时间捕获客户行为的时候，这尤其正确，因为不管时间序列正则与否，它们必然汇总用于客户特征标识。

数据也遭遇（使得我们同时遭遇）一些问题，如缺失值、不正确的数值和来自不同源的不一致的数据。一旦这类问题被确认，应该研究它们。最大的问题是未知的那些数据，即数据看似正确，但是由于某个原因实际上是错误的数据。

许多数据挖掘工作必须使用不太完美的数据。正如冒着蓝烟但仍然设法沿着街道跑的旧汽车一样，这些工作产生足够好的结果。如同爱尔兰剧作家Samuel Beckett所写的名剧*Wait for Godot*（中译名《等待戈多》）中的流浪汉，我们可以选择等待，直到完美来临，但那是干不成任何事情的方法；较好的选择是努力向前，不断学习，逐渐取得进步。

第 18 章　应用数据挖掘

你已经到达本书的最后一章，并且已经准备开始将数据挖掘用于公司业务。你确信，当数据挖掘已经融入公司的时候，整个企业将受益于对客户和市场与日俱增的了解、更集中关注的市场、销售资源的更有效利用，以及更多响应的客户支持。你也知道，理解在一本书中看到的东西与实际付诸实施之间有很大的差别。本章意在桥连这个隔阂。

由本书作者们创建的 Data Miners 咨询公司，已经帮助许多公司实施第一次的数据挖掘计划。虽然本章重点关注公司的第一次数据挖掘尝试，但真正关心的是如何增加数据挖掘计划成功的概率，而不管计划是第一个还是第五十个。本章集中前面几章的思想，并把这些思想应用于数据挖掘试验方案（pilot project）的设计。首先给出整合数据挖掘与企业的一般建议。然后讨论如何选择并且实现成功的试验方案。最后以一家公司的最初数据挖掘工作及其成功的事例作为总结。

18.1　开始

将数据挖掘完全整合到公司的客户关系管理（customer relationship management，CRM）策略，是一项巨大和令人畏惧的工程。沿着既定方法，采用可实现的目标和可测量的结果，逐渐地接近这个策略。最终目标是让数据挖掘很好地融入决策制定过程，使企业决策理所当然地使用准确、及时的客户信息。实现这个目标的第一步是，通过易处理的试验或概念验证（proof-of-concept）方案，产生可测量的投资回报，从而演示真正的数据挖掘的商业价值。应该选择本身有价值的试验，并且为企业案例提供一种坚实的基础，来证明在分析客户关系管理时做进一步投资是有价值的。

事实上，试验方案与任何其他数据挖掘计划没有什么不同。尽管有一些变化，但在试验方案中仍然描绘了数据挖掘良性循环（virtuous cycle）的所有四个阶段。概念验证被限定在预算和时帧之中。在试验方案中，通常需要修复的关于数据和程序的某些问题可能只出现在文件中。

提示： *在利用数据挖掘逐渐进行企业改革的工作中，试验方案迈出了有利的第一步。*

以下是我们与客户合作的一些数据挖掘试验方案实例中的主题语句：

- 及时找出 10 000 个最有可能在 10 月流失的高端移动电话客户，以便在九月开始举办电话营销活动。
- 在德克萨斯，参考即食的谷类食品，找出西班牙和非西班牙购物者的购买简档（profile）的差别，从而更好地指导西班牙语的广告营销活动。
- 通过发现最佳客户的共同点来指导扩充计划，并且定位于能发现相似客户的新市场。
- 在公司数据仓库的客户之间建立模型，识别市场研究（market research）片段，从而能够有针对性地将有关信息传递给适当的客户。
- 预测后几个月的债务回收预期程度，以便设法制定一个计划。

这些例子表明了数据挖掘致力解决的问题的多样性。在每种情况中，数据挖掘的挑战是找到并且分析适当的数据来解决商务问题。然而，这个过程首先从选择正确的示范方案开始。

18.1.1 从概念验证方案中能期待什么

当概念验证方案是完整的时候，下列各项是可用的：

- 原型模型开发系统（可能被外包，或者是生产系统的核心）
- 几种数据挖掘技术和工具的评估（除非预先确定工具）
- 一个修改商务过程和系统使之与数据挖掘一体化的计划
- 产生数据挖掘环境的描述
- 在数据挖掘和客户分析中的投资商务案例

即使已经决定投资数据挖掘时，概念验证方案仍是首次迈入数据挖掘良性循环的重要方法。沿着这个方法，应该会面临挑战和暂时的困难，因为这样一个方案涉及企业的几个不同部门，包括技术和操作部门，并且需要他们以也许不熟悉的方式一起工作。

18.1.2 识别概念验证方案

概念验证方案的目的是，当管理风险的时候，有效地发挥数据挖掘的效用。方案应该足够小，因而是实用的；足够重要，因而是有意义的。成功的数据挖掘概念验证方案将产生可测量结果的行为。为了寻找概念验证的候选者，研究现有的商务过程，来识别在哪些领域中数据挖掘可以提供结果能以美元测量的切实利益。也就是说，为进一步将数据挖掘整合到公司的营销、销售和客户－支持操作，概念验证应该创造可靠的商务案例。

吸引注意力和编制美元预算方案的一个好方法是，使用数据挖掘来满足真正的业务需求。最令人信服的概念验证方案重点关注已经被测量和分析评估的区域，并且在这些区域具有公认的进一步完善的需求。有可能的候选情况包括：

- 响应模型
- 默认风险模型
- 流失模型
- 使用模型
- 收益模型

在这些领域，改善预测准确度和改善收益之间有明确的联系。利用某些方案，容易对数据挖掘结果采取行动。这并不能说，重点关注日益增加的洞察力和理解，但没有任何与结果的直接联系，试验方案就不可能成功。然而，建立商务案例是比较困难的。

潜在的新信息用户时常具有创造性和丰富的想象力。在面谈期间，鼓励他们想象开发真实地学习客户关系的方法。同时，制作可用数据源的详细目录，识别（identifying）期望或必需的附加字段。在数据已经装入仓库的地方，学习数据字典和数据库模式。当源系统（source system）是操作系统的时候，研究未来提供数据的记录布局，并且开始了解那些熟悉系统如何处理和存储信息的人们。

概念验证选择过程的一部分工作是，对可用的记录和字段建立简档，以便初步了解数据中的关系，得到某些可能阻碍数据挖掘进程的数据问题的早期警告。这个工作可能需要一定量的数据清理、过滤和转换。

一旦确定了几个候选方案，就可以从以下几个方面来估计方案，包括依据结果采取行动的能力、潜在结果的有效性、数据的可用性和技术工作层面。关于每个试验方案，最重要的

问题之一是"结果将如何被使用?"如同在后面"成功的概念验证"部分中的例子阐明的一样，数据挖掘试验方案共同的命运是技术上的成功，但却未得到正确评价，因为没有人能领会利用这些结果做什么。

当然，也有许多源于IT的成功数据挖掘方案的例子。然而，当引导数据挖掘的人们没有定位在营销或者与客户直接交流的某个其他团体时，赞助或者至少来自这类团体的输入对于成功的方案是重要的。虽然数据挖掘需要与数据库和分析软件形成互动，但它主要不是IT方案，并且不应该尝试与所讨论的商务问题的拥有者隔绝。

提示：*数据挖掘试验方案可能基于公司内几个团体中的任何一个，但是，它必须总是包括团体中活跃的参与者，即所讨论的商业问题的拥有者。*

营销活动创造出好的概念验证方案，因为在大多数公司，已经有测量此类活动结果的文化。对照实验表明，与直接邮寄、电话推销或电子邮件活动相对应的统计意义显著的改进可以很容易地转化为经济利益。证明数据挖掘价值的最好方法是利用超越估计模型的示范方案，以模型为基础，实际测量活动的结果。尽管不可能，仍要仔细思考如何增加示范方案结果的经济价值。在有些情形中，测试根据历史数据从数据挖掘获得新模型就足够了。

成功的概念验证

数据挖掘概念验证方案可以说技术上是成功的，然而总的来说令人失望。例如，一家移动电话公司开始实施数据挖掘方案，以期更好地了解客户流失。该方案在识别几个高流失风险的客户片段中获得成功。利用被识别的团体，公司可以为留住这些客户提供激励。到目前为止，方案看起来是一个好的、能够返回可操作结果的概念验证。

这个数据挖掘模型发现一群高风险客户，由呼叫行为与其套餐计划不相匹配的用户组成。这些客户的一个子群位于低的月套餐计划，相应地，通话分钟数很少。此类计划对于不经常使用电话的人有意义，例如"安全用户"(safety user)，他们把电话留在汽车储藏柜中，很少打开，但是从紧急情况下电话可用这一点来说却是安全的。当这类用户改变打电话的习惯(如同有时候发生的那样，一旦他们认识到移动电话的有效性)的时候，他们最终不会使用比套餐计划更多的分钟数，因为需要为超出部分支付较高的每分钟费用。

由于被模型识别为高风险的群体被追踪，而且事实上他们确实陆续离开，公司宣布，数据挖掘方案是成功的。然而，因为发起数据挖掘计划的团体的主旨是探究新技术方案，而不是处理客户关系，所以没有采取任何行动。狭义地讲，方案的确是成功的。它证明数据挖掘能够识别高风险流失客户的概念。广义地讲，企业还没有为数据挖掘做好准备，因此，无法成功地对结果采取行动。

对于这些客户，企业面临着另外的挑战。只要匹配不当的客户继续保持原状态，支付昂贵的过度呼叫或特别昂贵的套餐计划，他们是相当有利可图的。把他们转移到省钱("正确规划"他们)的计划，可以非常好地减少流失，但也减少收益。哪一个更重要呢，是流失还是收益?通常，数据挖掘提出的问题和回答的问题一样多，而且某些问题的答案更多地取决于商业策略，而不是数据挖掘结果。

18.1.3 实现概念验证方案

一旦选择适当的商务问题，第二个步骤是，识别和收集能被转换为可操作信息的数据。数据源已经被确定为选择概念验证方案过程的组成部分。第三个步骤是，从那些源中提取数

据，并转变为前面章节中介绍的客户特征标识。在设计好的客户特征标识时，开始几次是不易处理的。这是资深数据挖掘者的帮助可能有价值的领域。

除了构造初始客户特征标识之外，还需要探索原型数据和模型开发环境。这个环境可能是由软件公司或数据挖掘顾问服务公司提供，或内部构建为试验方案的组成部分。数据挖掘环境可能由安装在专用的分析工作站上的数据挖掘软件系列构成。模型开发环境应该足够充分，以便允许进行多种数据挖掘技术测试。关于选择数据挖掘软件和建立数据挖掘环境，第16章已给出建议。概念验证方案的目标之一是，决定哪项技术对于解决特定的商务问题最有效。

使用原型数据挖掘系统涉及一个过程，即精炼在环境、现有的操作系统和决策支持计算环境之间的数据提取需求和接口。期望这个反复的过程能够提供所需的对未来数据挖掘环境的更好理解。早期的数据挖掘结果将提出新建模方法和对客户特征标识的改进。

一旦原型数据挖掘环境建立，使用它建立预言性模型（predictive model）。当概念验证方案已经明确，应用模型完成最初的识别高回报的任务。小心测量模型在历史数据上的性能。

没有在内部实际构建原型数据挖掘环境的情况下，通过使用外部设备，完成整个概念验证方案也是切实可行的。这种方法有利也有弊。从积极的方面看，数据挖掘顾问从在其他公司的数据上工作的经历获得丰富经验用来指导当前处理的问题。专家拥有应用广泛的数据挖掘工具和技术的知识和经验，而公司的任何一位职员是不太可能具有这种知识和经验的。从消极的方面看，如果顾问做了全部的实际数据挖掘工作，你和你的职员就不会获得更多的有关数据挖掘过程的知识。或许最佳的折衷方案是，组织同时包括外部顾问和公司内部人员的团队。

1. 基于发现采取行动

下一步是测量建模的结果。在某种情况下，最佳方法是使用历史数据（对于有益的比较，利用过时的例子更适宜）。另外的需要与其他团体更多合作的可能性是，建立对照实验，将基于数据挖掘采取的行动的效果与当前的基准线进行比较。这种对照实验对于已经有进行此类实验基础的公司尤其有价值。

最后，使用建模结果（不管来自历史测试或真实的实验）构建商务案例，将数据挖掘整合入稳定的商务运作中。

有时，试验方案的结果是得到对客户和市场的深入了解。在这种情况下，通过向商业人士提供深入了解，成功更多取决于主观性。虽然这似乎是较容易的概念验证方案，但是在数周之内发现结果，并给具有多年经验的商业人士留下深刻的印象，具有相当的挑战性。

许多数据挖掘概念验证方案没有雄伟目标，因为最初的设计是为了评估技术而并非应用的结果。最佳情况是，更好的模型和更好的商务结果之间的联系不再是假设，而是通过真实结果来证明。统计师和分析师可能对理论结果留下深刻的印象，而高级管理人员却不是这样。

用于显示在测试数据集上新模型取得的响应率提升度的图给人留下深刻的印象；然而，通过模型获得的新客户给人留下的印象更深刻。

2. 测量行动结果

测量数据挖掘模型本身的有效性和根据模型预测结果而采取的行动对商业的真实影响都很重要。

提升度是测量模型自身有效性的一种适当方法，提升度测量某个特定类型（例如响应者或默认者）的记录的集中程度随模型得分的变化。为了测量对商务的影响，需要更多的信

息。如果试验方案建立了一个响应模型，请继续追踪下列费用和收益：

- 开展活动和建立支持模型的固定费用是多少？
- 每位接受者获得的促销服务的费用是多少？
- 响应每一个促销服务的费用是多少？
- 积极响应的价值是多少？

最后一项似乎很明显，但通常被忽略。我们已经见到不止一项数据挖掘工作在开始后陷入困境，因为尽管结果表明，数据挖掘可以带来更多的客户，但是关于新客户的价值没有一个清晰的模型，因此对于获得的收益就没有清楚的了解。

设计好的营销测试细节超出了本书的范围，但控制数据挖掘模型的效能和使用服务或宣传信息的效能都很重要。这可以通过追踪四个不同群体的响应来完成：

- 群体 A，由数据挖掘模型选择用于接受促销服务。
- 群体 B，随机选择用于接受相同促销服务。
- 群体 C，也是随机选择但未获得促销服务。
- 群体 D，由模型选择用于接受促销服务，但事实上没有得到促销服务。

如果模型在发现适当的客户方面性能良好，群体 A 的响应率会显著地高于群体 B 的响应率。如果服务是有效的，群体 B 的响应率会超过群体 C。有时，模型在发现无效服务的响应者方面性能良好，在这种情形下，群体 A 和群体 D 有相似的响应率。每一轮两两比较，回答一个不同的问题，如图 18-1 所示。

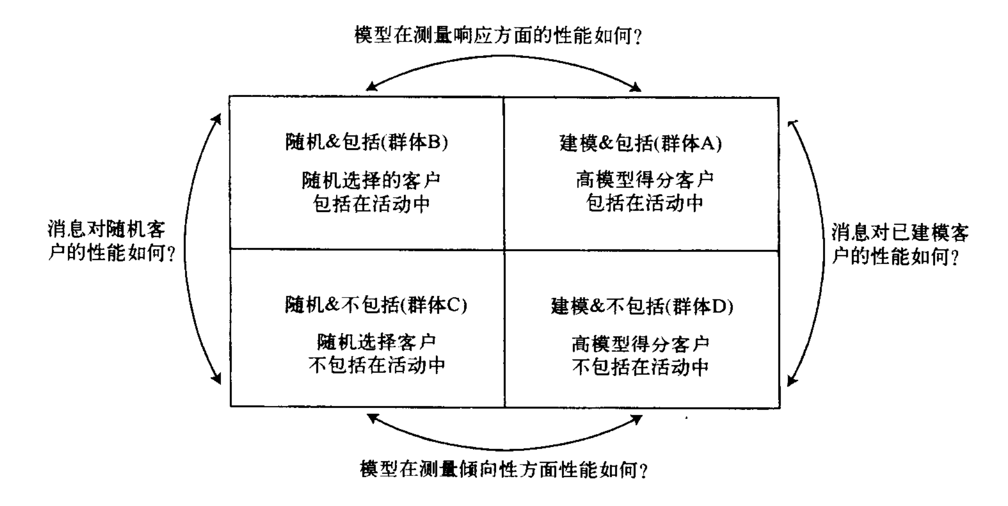


图 18-1　追踪四组不同的群体能够同时决定活动的效果和模型的效果

后一种情形确有发生。一家加拿大银行使用模型挑选那些可能通过直接邮寄活动开设投资账户的客户。事实上，通过模型挑选的人，不管是否收到了促销材料，开设投资账户的比例确实高于其他客户。在这种情况下，有一个简单的理由。银行使用有关投资账户的消息全方位地影响客户，像广告、分支机构的海报、宣传短片和客户打入电话并待机时播放的消息。相对于这些信息，直接邮寄宣传品是多余的。

18.2 选择数据挖掘技术

应用哪种或者哪些数据挖掘技术，取决于要完成的特定数据挖掘任务和分析可用的数据。在确定一项数据挖掘技术之前，首先把要处理的商务问题转化为一系列的数据挖掘任务，并且根据内容和数据字段的类型来领会可用数据的性质。

18.2.1 将商务目标转换为数据挖掘任务

第一步是，获得诸如"改善保持"的商务目标，并且将其转换成一项或多项第 1 章介绍的数据挖掘任务。回忆一下，本书所讨论的数据挖掘技术解决的六项基本任务是：

- 分类
- 估计
- 预测
- 关联分组
- 聚类
- 建立简档和描述

实现改善保持的商务目标的一种方法是，识别可能取消的用户（subscriber），找出取消的原因，制造一些类型的服务来解决他们关注的问题。成功的策略必须识别可能取消的用户，并且依照假设离开的理由将这些用户分配到一些群体。然后，为每个群体设计出适当的保持服务。

使用的模型集如果包含已经取消的客户实例和尚未取消的客户实例，那么本书讨论的许多数据挖掘技术可能把每位客户或多或少地标注为可能流失。识别独立的风险用户片段和理解每个组离开的动机的附加需求，提出了应该使用决策树和巧妙的衍生变量。

决策树的每片叶子有一个标签，在这种情况下，大概是"不可能流失"或"可能流失"。树上的每片叶子有不同的目标变量比例；这个流失者的比例可当作流失得分使用。每片叶子也有一组规则，描述在该处结束的客户。利用技巧和创造性，分析家也许能够将这些机械的规则变成可理解的离开的理由，一旦被理解，就可以采取行动来阻止。通常，决策树具有比为了开发特别服务和电话营销活动所期望的叶子更多。为了把这些叶子联合成更大的群体，把整个树枝作为群体，而不是单独的叶子。

值得注意的是，在这种情况下，选择决策树方法，是源于对了解流失（attrition）理由的欲望和区别对待子群的欲望。如果目标仅仅是最大限度地预测风险客户，而无需关心理由，就可以选择不同的方法。不同的商务目标需要使用不同的数据挖掘技术。如果目标是估计下个月每位用户使用的分钟数，神经网络（neural network）或者回归（regression）可能是较好的选择。如果目标是发现自然发生的客户片段，非定向聚类（undirected clustering）技术或简档和假设测试（hypothesis testing）是合适的选择。

18.2.2 决定数据的相关特性

一旦数据挖掘任务已经确定，并且用于缩小所考虑的数据挖掘方法的范围，可用数据的特征有助于更进一步地细化这一选择。用更一般的术语来说，目标是选择能够最小化数据转换的数量和难度的数据挖掘技术，这些转换是从数据得到有益的结果所必须执行的。

如同前一章的讨论，一定量的数据转换总是数据挖掘过程的组成部分。原始数据可能需要以各种不同的方式进行汇总，数据编码必须合理化，等等。不管选择的技术是什么，这些类型的转换是必需的。然而，对于有些数据挖掘技术，某些类型的数据引起一些特别的问题。

1. 数据类型

对于使用输入变量为数值型数值的数据挖掘技术来说，分类变量尤其有问题。能进行求和与乘法运算的这类数值型变量，迎合一些基于算术运算的数据挖掘技术的实力，例如回归、K平均聚类和神经网络。当数据具有许多分类变量的时候，决策树是相当有用的，尽管关联规则（association rule）和链接分析（link analysis）可能在某些情形中也是适用的。

2. 输入字段的数量

在定向数据挖掘应用中，应该有单一的目标字段或者依赖变量。其他的字段（不包括那些明确无关或者明显依赖于目标变量的字段）被当作潜在的模型输入。数据挖掘方法在成功地处理大量输入字段的能力方面是不一样的。对于特定的应用，这可能是在确定合适技术方面的一个因素。

一般来说，当字段的数目非常大的时候，依赖于调整权重向量的技术将陷入麻烦，其中的向量对每个输入字段分配一个元素。神经网络和基于存储的推理（memory-based reasoning）就是例子。关联规则面临不同的问题。该技术查看所有可能的输入组合；随着输入量的增长，在合理的时间内处理组合变得不可能。

决策树方法很少受大量字段的影响。构建树的时候，决策树算法识别在每个结点上贡献最多信息的单个字段，并且下一个规则片段仅仅以那个字段为基础。数十个或数百个其他的字段可能逢场作戏，但是在最终规则中不会表现出来，除非它们有助于解决问题。

提示：对于定向数据挖掘问题，当面临大量字段的时候，以构建决策树作为开始是一个好主意，即使最后的模型构建要使用不同的技术。决策树将会识别字段的一个好的子集，将其用作另一项技术的输入，该项输入可能淹没在最初的输入变量集中。

3. 自由形态文本

许多数据挖掘技术不能直接处理自由形态文本。但是很清楚的是，文本字段时常包含极有价值的信息。当分析独立的经销商向发动机制造商提交的维修声明时，机械工解释出错信息和修复问题方法的自由形态的记录至少与那些表示维修零件数目和所用工时的固定的字段同样有价值。

能处理自由文本的一种数据挖掘技术是基于存储的推理，即在第8章讨论过的最近邻方法之一。回忆一下，基于存储的推理基于测量数据库一条记录到所有其他记录的距离的能力，以得到相似记录的近邻。通常，发现适当的距离度量（measure）是一个使应用技术陷入困境的绊脚石，但是在信息检索领域的研究人员已经提出在两个文本块之间的好的距离度量。这些度量以文件之间的词汇重叠为基础，尤其是不常见的字和专有名词。网络搜寻引擎查找适当文章的能力是一个熟知的文本挖掘的例子。

如第8章所述，基于存储的自由形态文本推理也已经被应用于把工人按照产业和工作分类。这些工作分类基于美国人口普查的冗长表格所提供的书面工作描述，并且对新闻报道添加关键字。

18.2.3 考虑混合方法

有时，几种技术组合比任何单一方法的效果更好。这可能需要把单一数据挖掘任务分解

成两个或多个子任务。第 2 章的汽车营销例子是一个好的样本。研究人员发现，选择特定汽车型号的潜在顾客的最佳方法是，首先使用神经网络识别可能买汽车的人，然后使用决策树预测每位购车者会选择的特别型号。

另一个例子是，一家银行使用三个变量作为信用诱惑决策的输入。三个输入估计如下：

- 响应的可能性
- 来自该客户的第一年计划收益（revenue）
- 新客户未履行任务的风险

这些任务在几个方面显著不同，包括可能有用的训练数据（training data）的数量，看似重要的输入字段，以及检验预测的准确度所需要的时间长度。邮寄后不久，银行确切知道谁是响应者，因为诱惑计划包含一个最后期限，在这个期限之后的响应视为无效。在核对第一年的估计收益和实际数量之前，必须经过整整一年的时间，并且客户可能经过更长的时间才会“变差”。给定所有这些差异，并不令人惊讶的是，不同数据挖掘技术对每项工作可能都是最佳选择。

18.3 公司如何开展数据挖掘

多年来，作者见证了许多公司进行第一次数据挖掘的尝试。虽然每家公司的情形是独特的，但显现出一些共性。在每家公司，有一位负责数据挖掘方案的人，确实相信分析客户关系管理的力量和潜能的原因，通常是因为他或她已在其他公司见到这种事情。这位负责人通常不是技术专家，而且经常不做任何实际的技术工作。他或她的作用是作为组织人，建立数据挖掘团队和保护数据挖掘试验方案的赞助者的地位。

成功的努力越过企业边界，涉及营销和信息技术人员。团队经常是相当小的，时常只有4～5 个人，但仍然包括了解数据的人，了解数据挖掘技术的人和了解处理商务问题的人，并且至少一个人具有应用数据挖掘处理商务问题的经验。有时，这些角色中的几个角色可能汇集到一个人身上。

在所有的情况中，最初的数据挖掘试验方案解决了对企业来说真正至关重要的问题，在这种问题上，能够体现成功的价值。一些最佳试验方案是为测量数据挖掘的有用性而设计，方法是观察数据挖掘工作所建议的行动的结果。

其中一家无线电话服务提供商同意我们描述他的数据挖掘试验方案。

18.3.1 保持的对照实验

Comcast Cellular 公司是一家无线电话服务提供商，在 1996 年，该公司主要关注环费城附近三个州的区域，拥有 750 万人的市场。1999 年，Comcast Cellular 公司被 SBC 公司收购，现在已经是 Cingular 公司的组成部分，但是当这项试验研究进行的时候，它是一家地方服务提供商，面对快速增长的全国网络的激烈竞争。日益激烈的竞争意味着用户会面对很多竞争对手的服务，并且每个月都有很大比例的客户转向有竞争力的服务。正如行业所称，这种流失是困惑所在，因为即使新用户数轻易地超过离去者的数目，但获取一位新客户的代价时常在 \$500～\$600 的范围。流失已在第 4 章给出了详细的讨论。

面对更多竞争对手泰然自若地进入市场，Comcast Cellular 公司希望利用积极主动的工作热情，确保持续不断地抓紧现有的用户。困难在于了解哪些是风险客户，并且原因是什

么。对于任何保持活动，了解哪些是风险客户很重要，因为保持服务要花费公司的资金。向无论如何可能保留的客户提供诱导没有任何意义。理解什么动机使不同的客户片段离开是同等重要的，因为不同的保持服务适合不同的客户片段。提供免费的晚间和周末分钟数，对主要使用电话与朋友保持联络的客户可能非常有吸引力，但是商务用户的兴趣则不大。

试验方案是一个三方合作的关系，涉及 Comcast Cellular 公司、一群数据挖掘顾问（包括作者）和电话营销服务局。

- Comcast Cellular 公司按照自己的商务实践和程序，提供数据和专门技术。
- 数据挖掘顾问利用详细呼叫数据中使用模式（usage pattern），开发可能的背叛者的简档。
- 电话营销服务局与 Comcast Cellular 公司一起，使用简档开发拓展电话营销活动的保持服务。

该描述重点关注联合工作的数据挖掘方面。数据挖掘工作的目标就是识别一些群体，其中的用户在未来 60 天中有异常高的可能性会取消订阅。采用的数据挖掘工具使用类似决策树的规则归纳算法，创建由简单规则所描述的高风险客户的片段。在针对保留这些高风险客户的电话营销活动中，计划应该包括他们，使保持服务适合于通过数据挖掘发现的不同客户片段。实验设计允许比较三个群体：

- 群体 A 由模型判断为高风险的客户组成，对这些客户不进行任何干预。
- 群体 B 由模型判断为高风险的客户组成，对这些客户施以适当的干涉。
- 群体 C 代表普通客户人口。

研究设计如图 18-2 所示。当然，我们希望，与群体 B 和 C 相比，群体 A 的流失率高，从而证明模型和干预两者都是有效的。

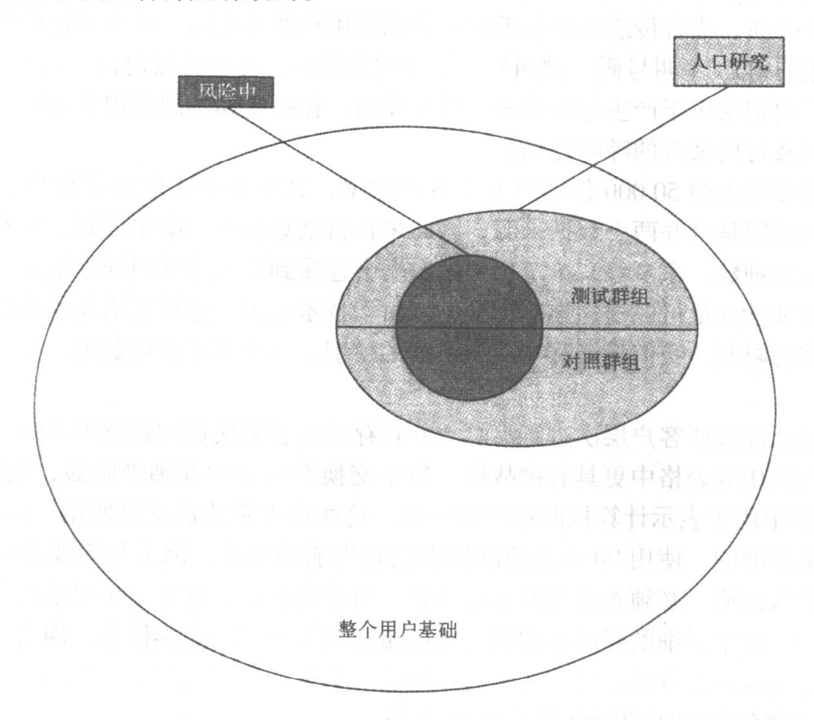


图 18-2 用于分析客户关系营销测试的研究设计

这里，方案遇到一点小麻烦。第一个困难是，虽然方案包括向被确认为可能取消的人进行电话营销呼叫的预算，但既没有预算也没有授权真正向被叫的人提供任何东西。另一个问题是呼叫中心的技术问题。为解决超出保持工作范围的特定问题（例如账单上的错误），将不满意的客户直接转移到电话公司的客户服务组是不可能的。然而另外的问题是，虽然客户数据库包括每个客户的家庭电话号码，但是事实证明，大约只有75%的电话号码是正确的。

最后，电话营销公司仅仅呼叫被测试的人和对照群组，询问一系列专门为得出他们的满意程度而设计的问题，并自愿谈及报告给客服中心的任何问题。不考虑这些毫无说服力的干预，测试群组60天的保持显然比对照群组好得多。很显然，仅显示公司特别关心客户的呼叫就足以减少流失了。

18.3.2 数据

在与客户的几次会谈期间，我们为试验的应用确认了两个数据来源。第一个来源是，由一家数据库营销公司建立的客户简档数据库。这个数据库包含每位用户的概要信息，包括套餐计划、电话类型、每月使用的本地呼叫分钟数、每月使用的漫游呼叫分钟数、到美国的某个移动电话市场的往返呼叫次数，以及许多其他类似的字段。

第二个来源是，无线交换机收集的呼叫明细数据。每次移动电话被打开，利用附近的发射站（cell site）开始双向交谈。发射站中转来自电话的数据，例如到中央交换局的序号和电话类型。交换局的计算机判断当前的电话应该打给哪一个发射站，并且将信息传回到电话，告诉它使用哪一个发射站以及调到什么频率。

当用户输入电话号码并且按下发送按钮的时候，号码被传递到中央交换机，它在正规线路上依次建立呼叫，或者传递给最接近另一位无线用户的发射站。每个交换产生详细的呼叫记录，包括用户ID、主叫号码、被叫号码、主叫发射站、呼叫持续时间、呼叫终止理由等。这些详细的呼叫记录用于产生每位客户的行为简档，包括诸如不同的被叫号码、每天某时的呼叫比例，以及每周某天的呼叫比例。

试验方案使用大约50 000位用户6个月的数据，其中有些人取消了账户，有些人没有取消。最初的意图是合并两个数据来源，因此来自营销数据库（账单计划、保有期、电话类型、使用的总分钟数、家乡等）给定用户的数据会连接到个人的呼叫详细记录。这样，基于两个数据源的独立变量可以建立惟一的模型。由于技术原因，这被证明是困难的，因此，由于时间和预算的限制，我们最终构建两个单独的模型，一个基于营销数据，另一个基于详细呼叫数据。

营销数据已经按照客户层次进行汇总，并且存储在容易接近的数据库系统中。把详细呼叫数据放到可使用的表格中更具有挑战性。每个交换有自己的卷轴带收藏，就像在20世纪60年代的电影中用于表示计算机的那些带一样。这些带不断地被反复使用，以至于90天的移动窗口总是当前的，使用90天之前的带来记录当前的呼叫。因为每天要录八卷音带，我们发现自己陷入困境，必须查看700多盘音带，每个音带必须逐个地手工装载到9轨道模拟音带驱动器中。由于详细的呼叫数据以加密的格式写入专门的交换设备，因此一旦装入，需要大量的预处理以便为分析做好准备。通过过滤那些与往返于流失模型人口的呼叫无关的记录，7000万条详细的呼叫记录降到了1000万条。

甚至在预言性建模开始之前，详细呼叫数据的简单分析也提出了许多可能逐渐增加收益的方法。一旦呼叫明细以可查询的形式备用，就可能回答如下问题：

- 进行许多短呼叫的用户要比那些只进行较少、较长呼叫的客户更忠诚还是更不忠诚？
- 呼叫失败导致呼叫客户服务吗？
- 对于移动电话到移动电话与移动电话到固定电话，用户的呼叫周期的规模是什么？
- 用户的使用如何分别按小时、月、工作日到周末而变化？
- 用户呼叫电台的热线吗？
- 用户呼叫语音邮件的频率是多少？
- 用户呼叫客户服务的频率是多少？

对这些问题和许多其他问题的回答表明，有一系列的营销活动用来刺激在特定时期以特定方式使用移动电话。此外，正如我们所希望的，围绕从呼叫明细构造的度量建立的变量，如呼叫周期大小，被证明可以极好地预测流失。

18.3.3 一些发现

数据挖掘隔离几个高风险流失的客户片段。其中某些片段比其他片段更具有可操作性。举例来说，事实证明，根据呼叫接入网络的位置来判断，经常往返于纽约的用户比往返于费城的用户更有可能流失。这是信号覆盖的问题。住在 Comcast 公司覆盖的区域并且经常往返于纽约的客户，发现自己大部分工作日的呼叫处于漫游（使用另一家公司的网络）状态。实际上，账单计划使得漫游的费用非常昂贵。经常往返于费城者的整个往返路程和工作日都保持在 Comcast 公司覆盖的区域内，因此不会产生任何漫游费用。因为变更覆盖区域和变更控制套餐计划的规则都不是研究的发起人所能驾驭的，所以这个问题不太好操作，尽管信息可能被其他企业所利用。

潜在地更具操作性的发现是，呼叫模式与套餐计划不匹配的客户处于高风险流失状态。存在两种情形，客户呼叫行为可能与套餐计划不适合。一个客户片段要支付比实际使用时间更多的分钟数。无线电话公司可以通过将他们吸引到较低的套餐计划，从而增加这些客户的终生价值。他们每个月的价值可能减少，但可能持续更长的时间。确认这一点的惟一方式是通过营销测试。毕竟，客户可能接受服务，每个月支付较少费用，但是仍然以相同的比率流失。或者说，流失的比率可能会降低，但不足以补偿近期的收益损失。

当用户签约不包括许多免费分钟数的低套餐计划时，发现自己时常使用的分钟数超出了计划所允许的时间，在呼叫行为和套餐计划之间发生了另一种错误匹配。由于额外的分钟收费比率高，这些客户终止支付比包含更多时间的更贵套餐计划更高的费用。将这些客户转移到较高的套餐计划可能会为他们节省一些费用，同时也增加来自他们每月账单的固定收益量。

18.3.4 实践出真知

Comcast 公司能够对结合数据挖掘和电话营销行动计划的直接成本/收益进行分析。根据这笔数据，Comcast 公司能够对将来的数据挖掘工作做出明智的投资决定。当然，事情在那里并没有真正结束；永远不会。

公司面对一系列全新的问题，这些问题基于来自初始研究的数据。新的假设已经形成和测试。电话营销工作的响应数据成为新一轮知识发现的素材。人们提炼出新的产品理念和服务计划。因为公司更了解客户，所以每个回合的数据挖掘从一个比较高的基础开始。那就是数据挖掘的良性循环。

18.4 小结

在商务环境中，成功引入数据挖掘需要使用数据挖掘技术解决真正的商务挑战。对于刚刚开始分析客户关系管理的公司来说，整合数据挖掘可能是件令人畏惧的工作。概念验证方案是开始的好方法。概念验证应该产生可靠的商务案例，进一步将数据挖掘整合到公司的营销、销售和客户 - 支持操作。这意味着方案应该是在一个区域中，在这个区域，容易将通过数据挖掘得到的改良的理解与改良的收益联系起来。

最成功的概念验证方案从定义明确的商务问题开始，并且使用与问题相关的数据产生行动计划。然后，以可控制的方式执行行动，并且仔细地分析结果，评估所采取的行动的效能。换句话说，概念验证应该包括数据挖掘良性循环的完整过程。如果这个初始方案是成功的，将会是众多之中的第一个。从整体而言，本章的主要小结也是本书的重要内容：只有应用于很有意义的问题时，数据挖掘技术才会成为有用的帮手。数据挖掘是一项需要技术专长的技术活动，但其成功与否由商务方面的效果来测量。

数据挖掘技术 市场营销、销售与客户关系管理领域应用（原书第2版）

本书是数据挖掘领域的经典著作，数年来畅销不衰。全书从技术和应用两个方面，全面、系统地介绍了数据挖掘的商业环境、数据挖掘技术及其在商业环境中的应用。自从1997年本书第1版出版以来，数据挖掘界发生了巨大的变化，其中的大部分核心算法仍然保持不变，但是算法嵌入的软件、应用算法的数据库以及用于解决的商业问题都有所演进。第2版展示如何利用基本的数据挖掘方法和技术，解决常见的商业问题。

本书涵盖核心的数据挖掘技术，包括：决策树、神经网络、协同过滤、关联规则、链接分析、聚类和生存分析等。此外，还提供了数据挖掘最佳实践、数据挖掘的最新进展和一些富有挑战性的研究课题，极具技术深度与广度。配套网站www.data-miners.com/companion提供了每章的练习和用于测试各种数据挖掘技术的数据。全书语句凝炼、清新，对复杂概念的实际应用进行了生动解释，是必不可少的数据挖掘教材。

作者简介

Michael J. A. Berry
Gordon S. Linoff

他们是专业的数据挖掘咨询公司Data Miners的创办人。他们合作出版了一些经典的数据挖掘著作，包括 *Data Mining Techniques*、*Mastering Data Mining* 和 *Mining the Web*（均由Wiley公司出版）。作为数据挖掘顾问，他们一起为北美洲、欧洲和亚洲的许多大公司提供专业咨询，把客户数据、呼叫数据、网络日志条目、销售点记录和账单文件变成有用的信息，用于改善客户体验。他们都有近20年在营销和客户关系管理方面应用数据挖掘技术的经验。



www.wiley.com



封面设计 · 陈子平



华章图书

上架指导：计算机/数据库

华章网站 http://www.hzbook.com
网上购书：www.china-pub.com
投稿热线：(010) 88379604
购书热线：(010) 68995259, 68995264
读者信箱：hzjsj@hzbook.com
ISBN 7-111-19056-4
定价：49.00 元

